UNIVERSITY OF CALIFORNIA, SAN DIEGO

Understanding disease through data driven biology

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Andrew Michael Gross

Committee in charge:

Professor Trey Ideker, Chair
Professor Kelly Frazer, Co-Chair
Professor Vineet Bafna
Professor Napolone Ferrara
Professor Clodagh O'Shea
Professor Richard Bruce Schwab

2016

The Dissertation of Andrew Michael Gross is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2016

DEDICATION

This work is dedicated to my family, friends, and colleagues who have somehow managed to put up with me throughout this entire PhD process.

EPIGRAPH

If you can't explain it simply,
you don't understand it well enough.

Albert Einstein

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Trey Ideker for his support as the chair of my committee.

Chapter 1, in full, is a reprint of “Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss” as it appears in *Nature Genetics* 2014. Andrew M Gross, Ryan K Orosco, John P Shen, Ann Marie Egloff, Hannah Carter, Matan Hofree, Michel Choueiri, Charles S Coffey, Scott M Lippman, D Neil Hayes, Ezra E Cohen, Jennifer R Grandis, Quyen T Nguyen, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of “Analysis of Matched Tumor and Normal Profiles Reveals Common Transcriptional and Epigenetic Signals Shared across Cancer Types” as it appears in *PLOS One*, 2015. Andrew M. Gross, Jason F. Kreisberg, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in Molecular Cell, 2016. “Methylome-wide analysis of chronic HIV infection reveals five-year increase in biological age and epigenetic targeting of HLA”. Andrew M. Gross, Philipp A. Jaeger, Jason F. Kreisberg, Katherine Licon, Kristen L. Jepsen, Mahdieh Khosroheidari, Brenda M. Morsey, Hui Shen, Ken Flagg, Daniel Chen, Kang Zhang, Howard S. Fox, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

VITA

2010 Bachelor of Science, University of Texas at Austin

2016 Research Assistant, University of California, San Diego

2016 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Gross, **A.M.,** Jaeger, P.A., Kreisberg, J.F., Licon, K. , Jepsen, K.L., Khosroheidari, M., Morsey, B.M., Shen, H., Flagg, K., Chen, D., Zhang, K., Fox, H.S., Ideker, T. Methylome-wide analysis of chronic HIV infection reveals five-year increase in biological age and epigenetic targeting of HLA. **In Revision.**

Gross, A.M., Kreisberg, J.F., and Ideker, T (2015). Analysis of Matched Tumor and Normal Profiles Reveals Common Transcriptional and Epigenetic Signals Shared across Cancer Types. **PLoS One** 10, e0142618.

Shen, J.P., Srivas, R., **Gross, A**., Li, J., Jaehnig, E.J., Sun, S.M., Bojorquez-Gomez, A., Licon, K., Sivaganesh, V., Xu, J.L, Klepper, K., Yeerna, H., Pekin, D., Qiu, C.P., van Attikum, H., Sobol, R.W., Ideker, T. (2015). **Oncotarget** 6, 35755-35769.

Gross, A.M. and Ideker, T. Putting Networks in Context. (2015) **Nature Biotechnology** 33, 720-721. (News and Views Article).

Choueiri, M.B., Shen, J.P., **Gross**, **A.**, Huang, J., Ideker T. and P. Fanta. (2015). ERCC1 and TS expression as prognostic and predictive biomarkers in metastatic colon cancer. **PLoS One** 10, e0126898.

Gross, A.M. and Cohen, E. E. (2015). Towards a Personalized Treatment of Head and Neck Cancer. **American Society of Clinical Oncology Educational Book**, 35. (Review/perspective article)

Raju, S.C., Hauff, S.J., Lemieux, A.J., Orosco, R.K., **Gross, A.M.**, Nguyen, L.T., Savariar, E., Moss, W., Whitney, M., Cohen, E.E., Lippman, S.M., Tsien, R.Y., Ideker, T., Advani, S.J., Nguyen, Q.T. (2015). Combined TP53 mutation/3p loss

correlates with decreased radiosensitivity and increased matrix-metalloproteinase activity in head and neck carcinoma. **Oral Oncology** 51, 470–475.

Hauff, S.J., Raju, S.C., Orosco, R.K., **Gross, A.M.**, Diaz-Perez, J.A., Savariar, E., Nashi, N., Hasselman, J., Whitney, M., Myers, J.N., Lippman, S.M., Tsien, R.Y., Ideker, T., Nguyen, Q.T. (2014). Matrix-Metalloproteinases in Head and Neck Carcinoma-Cancer Genome Atlas Analysis and Fluorescence Imaging in Mice. **Otolaryngology -- Head and Neck Surgery**.

Gross, A.M., Orosco, R.K., Shen, J.P., Egloff, A.M., Carter, H., Hofree, M., Choueiri, M., Coffey, C.S., Lippman, S.M., Hayes, D.N., Cohen, E.E., Grandis, J.R., Nguyen, Q.T. and Ideker, T. (2014). Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss. **Nature Genetics** 46, 939–943.

Hofree, M., Shen, J.P., Carter, H., **Gross, A.**, and Ideker, T. (2013). Network-based stratification of tumor mutations. **Nature Methods** 10, 1108–1115.

Wanda, P.A., Fine, M.S., Weeks, H.M., **Gross, A.M.**, Macy, J.L., and Thoroughman, K.A. (2013). Brevity of haptic force perturbations induces heightened adaptive sensitivity. **Exp Brain Res** 226, 407–420.

Gendron, J.M., Pruneda-Paz, J.L., Doherty, C.J., **Gross, A.M.**, Kang, S.E., and Kay, S.A. (2012). Arabidopsis circadian clock protein, TOC1, is a DNA-binding transcription factor. **Proceedings of the National Academy of Sciences** 109, 3167–3172.

Muralidhara, C., **Gross, A.M.**, Gutell, R.R., and Alter, O. (2011). Tensor Decomposition Reveals Concurrent Evolutionary Convergences and Divergences and Correlations with Structural Motifs in Ribosomal RNA. **PLoS ONE** 6, e18768.

FIELDS OF STUDY

Major Field: Bioinformatics

Studies in Cancer Genetics
Professor Trey Ideker

Studies in Circadian Rhythms
Professor Steve Kay

ABSTRACT OF THE DISSERTATION

Understanding disease through data driven biology

by

Andrew Michael Gross

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2016

Professor Trey Ideker, Chair
Professor Kelly Frazer, Co-Chair

Over the past five years, rapid technological advancement has allowed for a surge in data generation, allowing for unbiased genetic and epigenetic profiling of healthy and diseased individuals. Given these unprecedented advances in data generation, the key question arises of what we can learn from such rich datasets

that we didn't know via other means. Here, I take a data-first approach to tackle three different problems in which large data-collection efforts combined with novel analytic methods can shine new light on the biology of disease. In the first part, I exploit the comprehensive, multi-omics profiling provided by The Cancer Genome Atlas to conduct an analysis of the molecular and clinical features of head and neck squamous cell carcinoma (HNSCC) that govern patient survival. I find that among HNSCC tumors TP53 mutation is frequently accompanied by loss of chromosome 3p and that the combination of these events is associated with a surprising decrease in survival time. Continuing with analysis of TCGA samples, I then analyze a large pan-cancer set of patients with both tumor and adjacent normal tissue samples profiled. By observing shared transcriptomic and epigenetic changes across a large and diverse set of tumors, this analysis identifies those shared signals that are likely to be important for both the onset and progression of cancer cells. Finally I use genome-wide epigenetic profiles to develop and validate epigenetic models of human aging in whole blood and purified blood cells to quantify the impact of HIV infection on aging. This work finds that both chronic and recent HIV infection lead to an average aging advancement of 4.9 years, increasing expected mortality risk by 19%. Taken together these studies all explore new biological findings, while providing examples of the power data-driven analysis to aid in the understanding of biology and disease.

Chapter 1: Multi-tiered genomic analysis of head and neck cancer ties *TP53* mutation to 3p loss

Chapter 1.1: Abstract

Head and neck squamous cell carcinoma (HNSCC) is characterized by aggressive behavior with a propensity for metastasis and recurrence. Here we report a comprehensive analysis of the molecular and clinical features of HNSCC that govern patient survival. We find that *TP53* mutation is frequently accompanied by loss of chromosome 3p, and that the combination of both events associates with a surprising decrease in survival rates (1.9 years versus >5 years for *TP53* mutation alone). The *TP53*-3p interaction is specific to chromosome 3p, rather than a consequence of global genome instability, and validates in HNSCC and pan-cancer cohorts. In Human Papilloma Virus positive (HPV+) tumors, in which HPV inactivates *TP53*, 3p deletion is also common and associates with poor outcomes. The *TP53*-3p event is modified by mir-548k expression which decreases survival even further, while it is mutually exclusive with mutations to RAS signaling. Together, the identified markers underscore the molecular heterogeneity of HNSCC and enable a new multi-tiered classification of this disease.

Chapter 1.2: Introduction

It is increasingly appreciated that the diversity of clinical outcomes in HNSCC is likely a reflection of the molecular heterogeneity of the tumor population (Leemans *et al.*, 2011; Mroz *et al.*, 2013; Stransky *et al.*, 2011). Previous studies have led to the identification of a variety of genes and other molecular features for stratifying HNSCC tumors, such as efforts to cluster gene expression profiles to define subtypes (Chung *et al.*, 2004; Walter *et al.*, 2013, Pickering *et al.*, 2013; Temam *et al.*, 2007; Lui, *et al.*, 2013). To comprehensively define this heterogeneity of common tumor types including HNSCC, The Cancer Genome Atlas (TCGA) project has generated multi-tiered molecular profiles for over 7000 patient tumors, providing an unprecedented opportunity to study the complex interrelations among fundamentally different types of molecular events and clinical outcomes such as patient survival.

Here we have built on the infrastructure established by TCGA to systematically and transparently unravel these complex relationships for HNSCC. To this effect, we obtained all available molecular and clinical data from TCGA (Lawrence *et al.*, 2015) as of the January 15, 2014 Firehose run and have documented all data-processing and analysis in a series of IPython Notebooks (Perez and Granger, 2007) (**Methods**, **Supplementary Table 1.1**). Five tiers of data – somatic mutations, chromosomal aberrations, mRNA expression, microRNA expression, and clinical variables – were analyzed for a total of 378 HNSCC patients resulting in measurements of over 34,000 molecular or clinical values for each patient (**Supplementary Figure 1.1a**). Because old age and HPV

status are associated with distinct molecular profiles and clinical outcomes (Leemans *et al.*, 2011) (**Supplementary Figure 1.2**), we focused analysis on the 250 patients under 85 years of age with HPV+ tumors and complete molecular profiles.

Chapter 1.3: Results

Chapter 1.3.1: Identification of prognostic events in HNSCC

We first sought to distill this multi-tiered, genome-wide dataset into a set of informative molecular and clinical events with potential relevance to cancer. First, individual somatic mutations and mRNA expression levels were integrated with knowledge of human molecular pathways to define aggregate 'pathway-level events' (**Supplementary Figure 1.1b-e, Methods**). Second, both individual and pathway events were filtered to select those that occur at high frequency (somatic mutations, chromosomal aberrations) or differential expression (mRNA and microRNA levels) in tumor versus normal tissue. The result of this analysis was a pool of 878 total events combined over all five tiers of data (**Supplementary Figure 1.1a**).

Next, we screened for individual events within each data type that are strongly predictive of survival, identifying 82 prognostic events out of the 878 (**Figure 1.1a**, **Supplementary Table 1.2**). Among somatic mutation events, *TP53* mutation was most strongly predictive overall, resulting in poor prognosis (Hazard Ratio 2.9 ±0.8, Benjamini Hochberg corrected $P < 0.01$). As has been observed

previously, survival outcomes were dependent on the TP53 protein domain affected by the mutation or its predicted functional status (Poeta *et al.*, 2007) (**Figure 1.1b**). However, we found that patients with mutations predicted as non-disruptive of function nonetheless had worse prognosis than patients with wild-type *TP53* (Hazard Ratio 2.2±0.7, $P = 0.03$). Among copy-number alterations, the most significant survival association was with heterozygous chromosomal deletions on the 3p arm which also led to very poor prognosis (**Figure 1.1a**, Hazard Ratio 3.5±1.1, Benjamini Hochberg corrected $P = 0.002$). Further analysis of chromosome 3p revealed that many patients have a deletion spanning a large fraction of the arm with increasing frequency of deletion approaching a fragile site in the 3p14.2 region (Ohta *et al.*, 1996) (**Supplementary Figure 1.3**). Although general chromosomal instability (CIN) as well as deletion of many individual chromosomal regions have previously been implicated as diagnostic (Partridge *et al.*, 1996; Leemans *et al.*, 2011) and prognostic (Meredith *et al.*, 1995; Partridge *et al.*, 1996, Partridge *et al.*, 1999; Lui *et al.*, 2013) markers, we find that the 3p event in particular was responsible for the majority of the impact on survival when compared with global rates of gene deletion (**Figure 1.1c**).

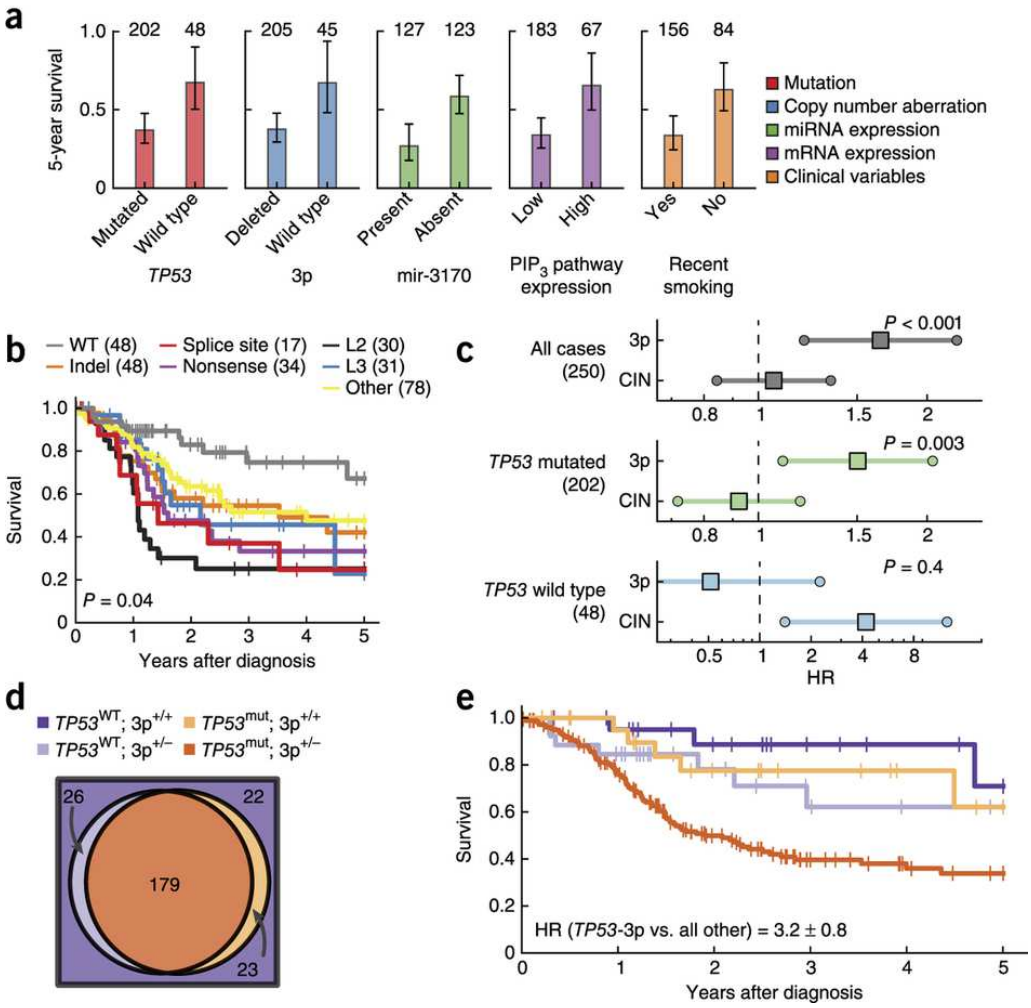


Figure 1.1: Prognostic effects and co-occurrence of TP53 and 3p. (a) Five-year survival (error bars indicate 95% CI) for the most significant events of each category (colors). Numbers above bars represent number of patients with each event. (b) Comparison of 5-year survival for patients with different types of non-silent TP53 mutations verses wild-type patients. L2 and L3 represent TP53 binding domains. Numbers in parentheses represent number of patients with a given mutation, patients with multiple TP53 mutations are represented multiple times in this plot. P-value represents log-rank test for TP53 mutation types excluding wild type. (c) Hazard ratios for multivariate Cox model fit with 3p deletion and global deletion rate (CIN) across different patient sets (age covariate not shown, error bars indicate 95% CI, p-values represent significance of likelihood ratio test for model fit with and without 3p deletion). (d) Venn diagram showing co-occurrence of TP53 mutation and deletions on the 3p chromosome. (e) Kaplan-Meyer curves showing survival outcome for all combinations of 3p deletion and TP53 mutation events (colors correspond to patient subsets in panel d).

Chapter 1.3.2: TP53 and 3p events co-occur and their combination predicts worse clinical outcome

It has previously been shown that genetic alterations often act by redundant or synergistic mechanisms to confer a growth advantage in the tumor (Cirello *et al.*, 2012; Bredel *et al.*, 2009). Under the hypothesis that individual events might act in concert, we next examined the 82 prognostic events for pairwise association across the patient cohort. This analysis identified 33 pairs of events that were significantly co-occurring or mutually exclusive (**Supplementary Table 1.3**). Among these, a particularly striking finding was that mutation of *TP53* and deletion of 3p occur very frequently together, in 179 of 250 HPV– tumors (**Table 1.1, Figure 1.1d**). While mutation of *TP53* has previously been associated with chromosomal instability (Leemans *et al.*, 2011), we found that *TP53* mutation associates with 3p loss far more frequently than it does with deletions in other chromosomal regions (**Supplementary Figure 1.4, Supplementary Tables 1.4-1.6**). Moreover, the combination of *TP53* and 3p events led to significantly worse survival than was predicted by either event independently or additively. Thus the synergistic interaction between *TP53* and 3p, with respect to both co-occurrence and survival, supports a clear molecular stratification of HNSCC tumors with and without this combination of events (**Figure 1.1c-e, Methods, Supplementary Figure 1.5, Supplementary Table 1.7**).

We found that the *TP53*-3p combination of events is associated with advanced tumor stage, although the stratification remains prognostic at all stages (**Supplementary Figure 1.6**). Furthermore, the prognostic effect cannot be

explained by clinical covariates alone and is particularly strong for smokers under 75 years old (175 patients, the majority of the TCGA cohort) for which the hazard ratio was 5.1 for the *TP53*-3p event relative to patients without this combination (**Supplementary Figure 1.7, Methods**).

To explore whether the interaction between *TP53* mutation and 3p deletion could be replicated in new patients, we obtained 126 additional HNSCC HPV– samples that had been deposited in TCGA while our initial study was underway (not included in the January 15, 2014 Firehose run). While these new patients did not yet have sufficient clinical follow-up for survival analysis, we indeed observed the same high co-occurrence of *TP53* mutation and 3p deletion (**Table 1.1**).

We also analyzed clinical follow-up data for 48 HNSCC HPV– tumors from the University of Pittsburgh Medical Center (Stansky *et al.*, 2011) for which the exome sequencing and copy number profiles had been previously collected after surgery (UPMC cohort, **Supplementary Table 1.1**). We observed that in this cohort, patients whose tumors contain the *TP53*-3p aggregate event have substantially worse prognosis than patients with *TP53* mutation alone, confirming the very large effect seen in the TCGA population (**Figure 1.2a** and **Table 1.1**). *TP53* and 3p events also co-occurred in the UPMC cohort, although with a lower effect size than in the two TCGA cohorts (**Table 1.1**); we suspect this is due to the much higher error rate of DNA sequencing in the earlier UPMC study, resulting in false-negative mutation calls (**Methods**).

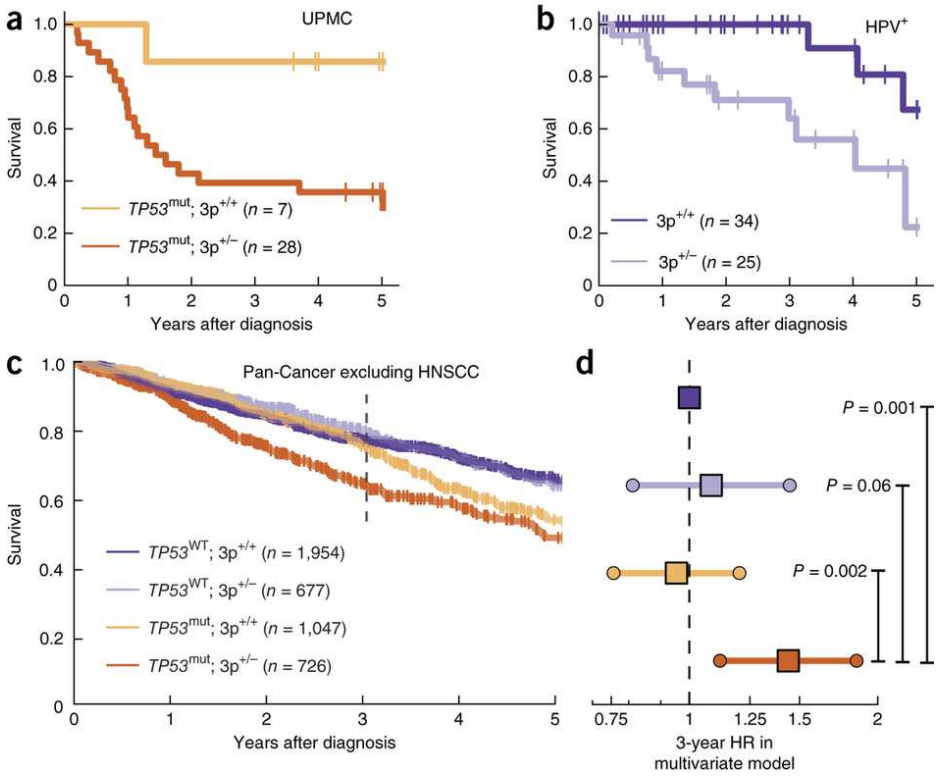


Figure 1.2: Replication of TP53-3p association. (a) Survival comparison of patients with TP53-3p aggregate event versus those with only TP53 mutation in the independent UPMC cohort. (b) Loss of 3p chromosomal arm is associated with lower survival in patients with HPV+ tumors (TCGA and independent cohorts). (c) Assessment of 3p loss and TP53 mutation association in TCGA Pan-Cancer cohort (HNSCC excluded). (d) Corresponding hazard ratio for multivariate model of three-year truncated survival (shown by dotted line in panel (c) when controlling for tissue type, age, and stage covariates. Error bars indicate 95% confidence.

We also sought evidence for the *TP53*-3p combination in patients with HPV+ tumors, in which TP53 is inactivated via interaction with HPV viral protein (Thomas *et al.*, 1999; Marur, *et al.*, 2010). Analysis of 59 HPV+ tumors from the TCGA and UPMC cohorts showed that *TP53* mutation is very rare in the presence of HPV (Odds Ratio 0.01, $P = 10^{-27}$ by Fisher's Exact Test), consistent with the expectation that the mutation confers little selective advantage once TP53 is inactivated by HPV. Among HPV+ tumors, the 25 tumors with 3p deletion had significantly worse prognosis than the 34 without the 3p event (Hazard Ratio 5.5 ±

2.6, P = 0.004). This finding lends further support for interaction between *TP53* and chromosome 3p with respect to survival and stratifies the growing population of patients with HPV+ tumors (Marur, *et al.*, 2010) (**Figure 1.2b**).

Another question was whether the *TP53*-3p interaction is specific to HNSCC or has broader support across diverse tissues. For this purpose, we performed a pan-cancer analysis based on all publicly available molecular data in TCGA (excluding HNSCC patients), covering 4404 patients over an additional 17 cancer types (Cancer Genome Atlas Research Network, 2013) (**Methods**). Although these tissues are molecularly heterogeneous and present with different patient outcomes (**Supplementary Figure 1.8a-c**), we nonetheless found compelling evidence for both the co-occurrence and impact on survival of *TP53* mutation and 3p deletion in this broader cohort, even when tissue type, patient age, and staging are accounted for (**Figure 1.2c-d**, **Table 1.1**).

Table 1.1: Co-occurrence and survival interaction of TP53 and 3p events.

			Co-occurrence of *TP53* / 3p events		Survival Interaction *TP53*-3p versus *TP53*	
Cohort		*n*	Odds Ratio	*p*	Hazard Ratio**	*p***
TCGA	Discovery	250	6.6	10^{-4}*	5.6	0.001
Recent TCGA	Validation	126	10	10^{-6}	ND	ND
UPMC	Validation	48	2.5	0.2	6.3	0.01
Pan Cancer	Validation	4404	2.0	10^{-25}	1.4	0.002

* Bonferroni corrected for test space

** Univariate model in patients under 75 years of age only

Chapter 1.3.3: Characterization of subtypes defined by combined *TP53*-3p event

Finally, we investigated whether the major subtypes defined by *TP53* and 3p status (**Figure 1.1e**) could be subdivided further by additional molecular markers (**Methods**). Indeed, we found that the 179 patients with the combined *TP53*-3p event were well stratified by the additional presence of microRNA mir-548k (**Figure 1.3a, Supplementary Figure 1.7c**) or mutation of the *MUC5B* gene (**Figure 1.3b, Supplementary Figure 1.7d**), both of which were associated with worse prognosis. Mir-548k is near *CCND1* and *FADD* on 11q13.3, which is commonly amplified in HNSCC (Meredith, *et al.*, 1995). Very recently, this micro-RNA has been shown to have oncogenic behavior in Oesophegeal Squamous Cell Carcinoma cell lines (Song *et al.*, 2014). While we found that 11q13.3 amplification

is associated with survival to a lesser degree than mir-548k expression, the prognostic effect seems to be specific to the expression of the micro-RNA (**Figure 1.3c**, **Supplementary Figure 1.9**).

Among patients lacking the *TP53*-3p event combination, we found strong enrichment for mutations to Caspase 8 as well as Ras and components of Ras signaling (**Table 1.2**, **Supplementary Figure 1.1b**). These enrichments were replicated in the TCGA molecular validation cohort (**Table 1.2**). The mutual exclusivity of Caspase 8 or Ras with *TP53*-3p provides further support for a *TP53*-3p defined subtype, and it implicates alternative routes to tumor progression in the absence of the *TP53*-3p event.

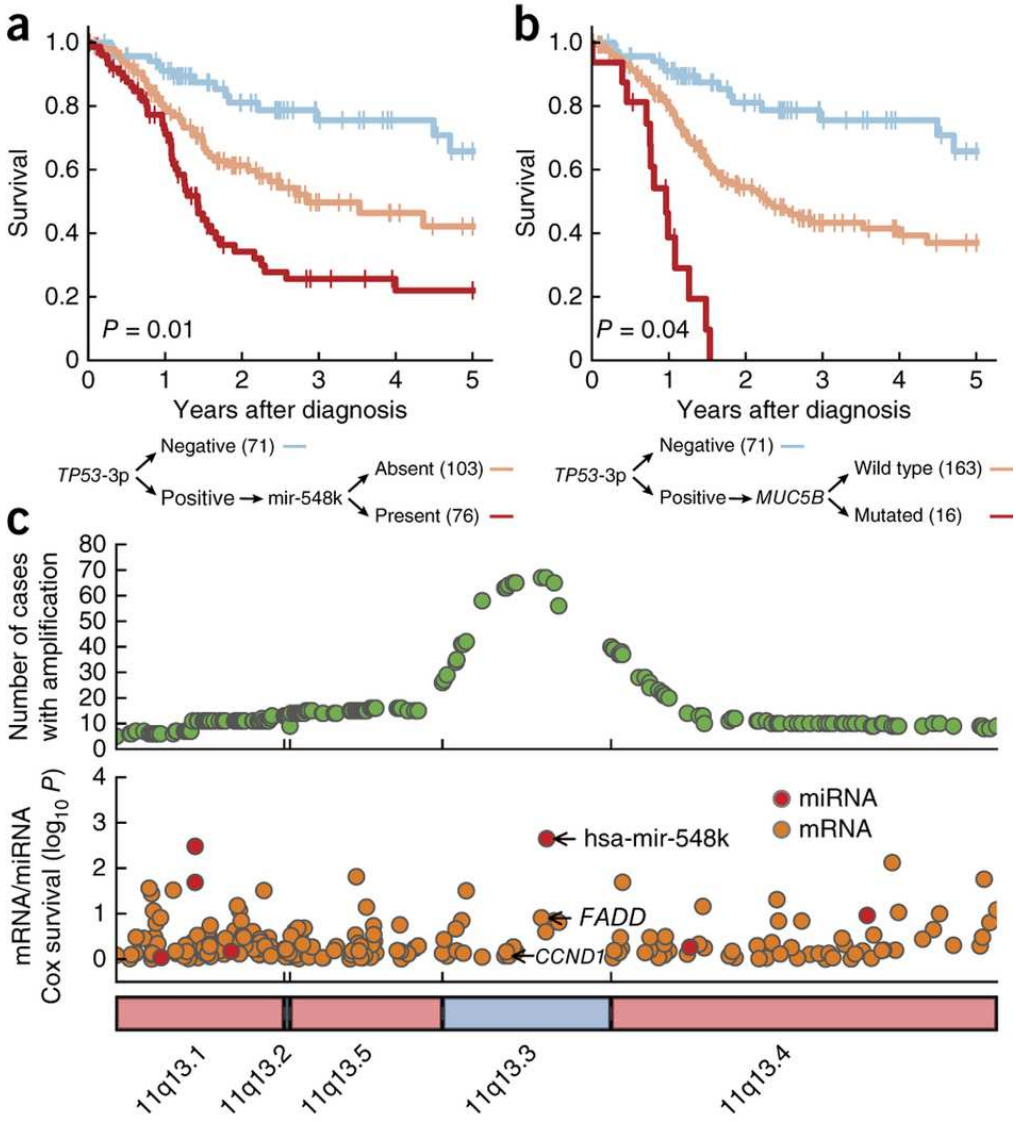


Figure 1.3: Characterization of molecular subtypes defined by the TP53-3p aggregate event. Patients with the TP53-3p aggregate event can be further stratified by the presence of a, mir-548k or b, MUC5B. (c) Frequency of high gain amplification (top panel) and association with patient survival for gene / miRNA expression (bottom panel) along the 11q13 chromosomal segment. P-values in a and b are Benjamini-Hochberg-corrected for 1008 events a secondary prognostic biomarker screen (Methods). All survival associations are calculated by a likelihood ratio test with age and year of diagnosis used as covariates in the set of 179 patients with the TP53-3p event (TP53-3p negative curves shown for comparison, but not used in computation).

Table 1.2: Co-occurrence of TP53-3p aggregate event and gene mutations.

			Co-occurrence of *TP53*-3p event and *CASP8* mutation			Co-occurrence of *TP53*-3p event and *RAS* Signaling Pathway† mutation		
Cohort		*n*	*# patients mutated*	Odds Ratio	*p*	*# patients mutated*	Odds Ratio	*p*
TCGA	Discovery	250	21	0.13	3×10^{-3}*	23	0.11	3×10^{-4}*
TP53-3p positive		179	6			6		
TP53-3p negative		71	15			17		
Recent TCGA	Validation	126	20	0.038	7×10^{-8}	21	0.86	5×10^{-6}
TP53-3p positive		81	2			4		
TP53-3p negative		45	18			17		

† Biocarta SOS1 Mediated RAS Signaling Pathway (Reacome 524)

* Bonferroni corrected for test space of 121 gene and pathway mutation events

Chapter 1.4: Discussion

As we approach a full inventory of driver events in cancer (Lawrence *et al.*, 2014), a key next step is to map and decode the complex network of interactions among individual events. Here, such an analysis was performed to identify a definitive stratification of head and neck cancer based on the largest tissue bank and dataset in existence. We have shown that *TP53* mutation, a well-studied driver event which leads to poor patient survival, is nearly always accompanied by specific loss of chromosome 3p (**Figure 1.1d**, **Table 1.1**). As has been argued for other cancer mutations (Bredel *et al.*, 2009; Xing *et al.*, 2012), the frequent co-occurrence of *TP53* and 3p alteration implies a selective advantage of cells acquiring both genomic events. In this study, the detection of the *TP53*-3p interaction was possible due to the high prevalence of each event individually, and their high (marginal) associations with patient survival.

While our study focused almost entirely on a single compelling interaction, our full analysis uncovered an additional 32 interactions in HNSCC which remain to be investigated (**Supplementary Table 1.3**). It is likely that this number is an underestimate, as low frequency and/or non-prognostic events were not evaluated. As cancer cohorts become larger, analyses such as this will become more powered, creating the opportunity to re-evaluate the cancer landscape from the perspective of pairwise and ultimately higher-order interactions among events.

Our analysis identifies two distinct clinical and molecular paths to cancer in HPV+ HNSCC patients. The first group, characterized by *TP53* mutation and loss of the 3p chromosome, is associated with advanced clinical stage and common

risk factors such as smoking. Nonetheless, this group tends to have very poor outcomes even when evaluated independently of these risk factors (**Supplementary Figure 1.7**). The second group of patients, lacking the *TP53*-3p combination of events, is characterized by mutations to RAS signaling and Caspase 8 (**Table 1.2**) and, ultimately, less aggressive tumors.

Further study is clearly warranted to elucidate the molecular underpinnings of these two groups of patients, with the goal of using such molecular stratification alongside clinical variables to inform patient treatment. Open questions relate to mechanism and the ordering of *TP53* and 3p events. What is the factor or factors encoded on chromosome 3p that are responsible for the interaction with *TP53*? Does one event necessarily precede the other and is a particular order required for poor survival? It is plausible that genomic instability primed by *TP53* mutation gives rise to loss of activity of a key factor encoded on chromosome 3p, but other scenarios are possible. Regardless, since the interaction of 3p with *TP53* or HPV status is independent of tumor stage, treatment of HNSCC patients might be modified to coincide with this specific molecular classification. In HPV+ HNSCC, the need for patient-tailored treatment programs is especially great, as we are currently in an era where we have maximized toxicity of existing regimens without necessarily improving outcome in cancers.

Our results also underscore the importance and value of public efforts such as TCGA in gathering, organizing, and distributing genomic data. Our work builds on the exemplary TCGA data collection and analysis pipeline (Cancer Genome Atlas Research Network, 2013) to integrate data across different measurement

platforms, with the goal of finding higher-order interactions of molecular events. Following the example of TCGA, we have documented and made public all analyses conducted in this study, ranging from data download to processing, exploratory analyses, statistical modeling, and visualization (**Methods**). With such a large and complex dataset, transparency and reproducibility of analysis is essential to provide a clear understanding of the methodology and to allow for further mining of results and extension to new datasets.

Chapter 1.5: Methods

Chapter 1.5.1: Availability

All data-retrieval and processing steps are documented in a series of IPython notebooks (Perez and Granger, 2007) available along with source code online at (https://github.com/theandygross/TCGA). These notebooks provide fully executable instructions for reproduction of the analyses and generation of figures and statistics for this study.

Chapter 1.5.2: Molecular Data

Data were obtained from The Cancer Genome Atlas Genome Data Analysis Center (GDAC) Firehose website (https://confluence.broadinstitute.org/display/GDAC/) using the firehose_get data-retrieval utility. All data were downloaded from the January 15th, 2014 standard data and analyses run unless otherwise specified. In order to maintain coherency of the analysis across different data layers and cancer types, we used Level 3

normalized molecular data as the input to our analysis. The use of the GDAC pipeline is intended to make these results easy to update as more TCGA data become available.

For a number of pan-cancer samples we generated mutation calls from TCGA aligned BAM files obtained from the UCSC Cancer Genomics Hub (https://cghub.ucsc.edu/). These calls were only used for patients with sequenced exome data that have yet go through the Firehose processing pipeline. Somatic mutation calls were made by running the MuTect mutation calling program (Cibulskis *et al.*, 2013) and the Genome Analysis Toolkit (GATK) SomaticIndelDetector (McKenna *et al.*, 2010) function on targeted regions with default parameters. All steps for downloading and processing this data are documented in the analysis notebooks and accompanying software repository. All mutation calls generated for this analysis are included as **Supplementary Table 1.8**. While these calls have yet to go through manual curation, we benchmarked this pipeline against TCGA working group mutation calls and found very high overlap with 94% sensitivity and 96% specificity.

Chapter 1.5.3: Pathway Data

Pathway data were downloaded from the Molecular Signatures Database (mSigDB) (Liberzon *et al.*, 2011). Version 3 of the canonical pathway gene-sets was used for this analysis.

Chapter 1.5.4: Candidate Biomarker Construction

Mutation calls were extracted from the annotated MAF files obtained from the Firehose and filtered to include only non-silent mutations. Each patient was

associated with a binary vector in which each position represents a gene; the position is set to 1 if the gene is observed to harbor one or more mutations in the patient and set to 0 otherwise. Mutation meta-markers were constructed by collapsing genes within a pathway gene-set via a logical OR such that the pathway is considered altered in a patient if any of its genes have a mutation (**Supplementary Figure 1.1b-c**). Pathway markers that were characterized by a single highly mutated gene or were highly correlated with mutation rate (Mann-Whitney U test, $P < .01$) were filtered.

Copy-number aberrations were extracted from the GISTIC2 (Mermel *et al.*, 2011) processing pipeline included in the standard Firehose analysis run. For biomarker construction data aggregated on significantly altered lesions (as deemed significant at the default 99% confidence settings) were used.

mRNA and miRNA expression data were obtained from the Level 3 normalized gene-by-patient matrices generated as part of the Firehose analysis pipeline. Data were $\log_2$ transformed. Genes/ miRNAs were first filtered based on differential expression comparing the full set of tumor expression profiles with the 34 profiles available for matched normal tissue (t-test, cutoff at $P < .01$). A pattern of background expression was estimated by taking the first principal component of non-differentially expressed genes or miRNAs. This background signal is meant to approximate the most common non-tumor related variation in expression due to inherent properties of the cohort such as population substructure or tissue specific expression changes. Real valued features with high correlation (Pearson Correlation, $P < 10^{-5}$) to this background expression pattern were filtered.

For the survival analysis, only the top 300 (of a possible 20502) differentially expressed genes were included in the analysis to limit the burden of multiple hypothesis correction (all 251 differentially expressed miRNA were used).

Markers used in this analysis consisted of binary markers and continuous valued markers. Binary markers were used when expression was only present (having more than ½ read per million) in a moderate fraction of the cohort (between 20 patients and half of the cohort). Real valued gene and miRNA expression levels were used for differentially expressed features not assigned as binary markers. Gene expression meta-markers were constructed from the loading of the first principal component of the reduced gene-by-patient matrix defined by each gene set. Due to similarity of gene-sets causing redundant gene expression meta-markers, marker pairs with high correlation (Spearman rho > .7) were reduced to a single informative marker by choosing the marker with the greatest differential expression. For the survival analysis, continuous valued markers were transformed into binary events prior to testing by setting a threshold that minimized the difference in variance between the resulting two groups. This was used to capture the skew of the distribution and assign the patients on the tail of the expression distribution as having an expression event (**Supplementary Figure 1.1e**).

Chapter 1.5.5: Clinical Data

Clinical data were downloaded directly from the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/). All outcomes reported relate to all-cause survival. Survival times were censored after five years to reduce the confounding

effect of patient age. For **Figure 1.2d**, survival times were censored after three years to show the specific effect within this time window, but all other figures and all statistics cited in the paper use five-year survival. While data on co-morbidity is limited for this cohort, from other studies we can estimate the competing mortality within this time-frame to be about 20% (Mell *et al.*, 2010; Farshadpour *et al.*, 2011). We expect the actual effect of such confounding to be minimal as separation in the survival curves that we observe generally occurs within the first two years, during which time we expect non-cancer associated death rates to be much lower.

For the primary and secondary survival screens, clinical data with missing data were used but statistics were only calculated on patients with data reported. In multivariate analysis (**Supplementary Figure 1.7**) missing value indicators were used.

Chapter 1.5.6: HPV Status

HPV calls from sequencing data were obtained from the TCGA HNSCC analysis working group. Due to the incompleteness of this dataset, this information was supplemented with HPV status called from a PCR-based MassArray Assay diagnostic provided on the TCGA data portal for patients where sequence-based data were not available.

Chapter 1.5.7: Prioritization of Prognostic Events

Feature selection is preformed prior to prognostic event prioritization. Events are selected for which at least 5% of patients are assigned to each group.

Prognostic events (**Figure 1.1a**) are prioritized via a likelihood ratio test comparing a Cox-proportional hazards model (Cox, 1984) fit with a candidate biomarker and covariates against a null model fit with the covariates alone. Age and the binary variable patient age > 75 are used as covariates (both age variables are used to model a non-linear association of patient age with survival). A multiple-hypothesis testing correction is employed which uses the method of Benjamini and Hochberg (1995) to control for the false discovery rate across the entire pooled space of tested features. After multivariate testing, a univariate log-rank test is assessed for each event and features with high multivariate significance, but low univariate significance ($P < 0.05$) are filtered from the pool of prognostic events.

As discussed in the text and in **Figure 1.3**, we conducted a second prognostic screen within the 179 patients with the *TP53*-3p aggregate event. For this analysis feature construction was repeated, resulting in 1008 candidate biomarkers (note that this number was higher than the primary screen due to more events passing the 5% threshold). During this secondary screen, we found the patient year of diagnosis to have a large impact on outcomes. For this reason we included this variable as a covariate in this screen.

Chapter 1.5.8: Statistical Analysis of TP53-3p Interaction on Survival

To assess the role of an interaction term in a statistical model of patient outcomes we performed leave-one-out cross-validation on a logistic regression model as shown in **Supplementary Figure 1.5**. To convert the survival data into a binary classification problem, we organized patients into two classes depending on whether they were surviving or deceased at T years after surgery. In this

analysis, the ratio of deceased to surviving patients is artificially high due to the ability to observe a death in a shorter followup than the full time interval required to annotate a patient as surviving (i.e. the basis of the Cox censorship problem). To reduce this bias, we removed patients with an observed death but a time of surgery after a set year (2013 – (T – 1)). As the problem was often unbalanced (the number of surviving patients differed from the number of deceased), re-weighting was preformed to give both classes equal weight. A multivariate Cox model fit to the most significant model is also shown in **Supplementary Table 1.7**.

Chapter 1.5.9: Pan-cancer Analysis

Pan-cancer data were downloaded and processed in the same manner as the HNSCC cohort. 3p chromosomal status was estimated via the median copy number of the twelve genes on the 3p14.2 locus.

In order to limit the heterogeneity of the pan-cancer cohort such that differences in molecular characteristics could be assessed, we performed a number of pre-processing steps. This reduced the patient cohort from 7081 to 4404 patients appropriate for survival analysis through the following filters:

- Only primary tumors were used for all patients, metastatic tumors were discarded.
- Glioblastoma patients were excluded due to the extremely low survival rate (6% five year survival).
- Diffuse large b-cell lymphoma, kidney chromophobe, thyroid carcinoma, and prostate adenocarcinoma patients were removed due to extremely

high rates of survival in the cohorts (84%, 86%, 90%, and 96% five year survival).

- Adrenocortical carcinoma, esophageal carcinoma, and pancreatic adenocarcinoma were excluded due to low sample counts (14, 39, and 69 patients in each tissue, respectively).
- Patients older than 85 years of age were excluded from the analysis to limit confounding from age (115 patients, Hazard ratio = 2.2 ± 3).
- Patients with high levels of residual tumor were excluded (66 patients, Hazard ratio = 2.9 ± .5).
- Stage IV patients were excluded (612 patients, Hazard ratio = 2.0 +/- .1)
- To limit circularity, HNSCC patients were excluded from all pan-cancer calculations but remain **Supplementary Figure 1.8** to allow for comparison to other tissue types.

Chapter 1.6: Contributions

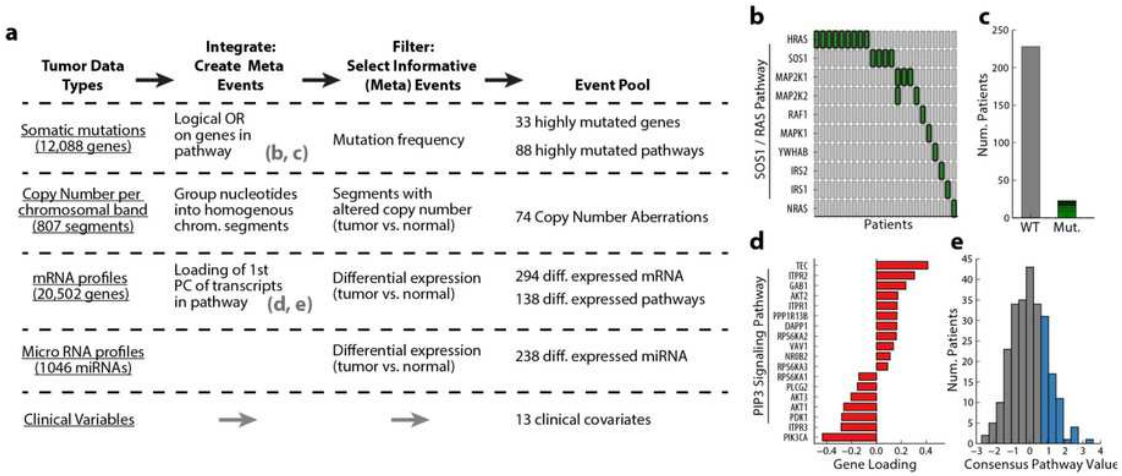
A.M.G., R.K.O., and T.I. conceived the study. A.M.G carried out most analyses. R.K.O., J.P.S., M.C., C.S.C, E.E.C., S.M.L, Q.T.N., and D.N.H. provided expertise. M.H. and H.C. aided in bioinformatic analysis. A.M.E. and J.G. collected and compiled clinical follow-up data for UPMC cohort. A.M.G. and T.I. wrote the manuscript with assistance from other authors.

Chapter 1.7: Acknowledgements

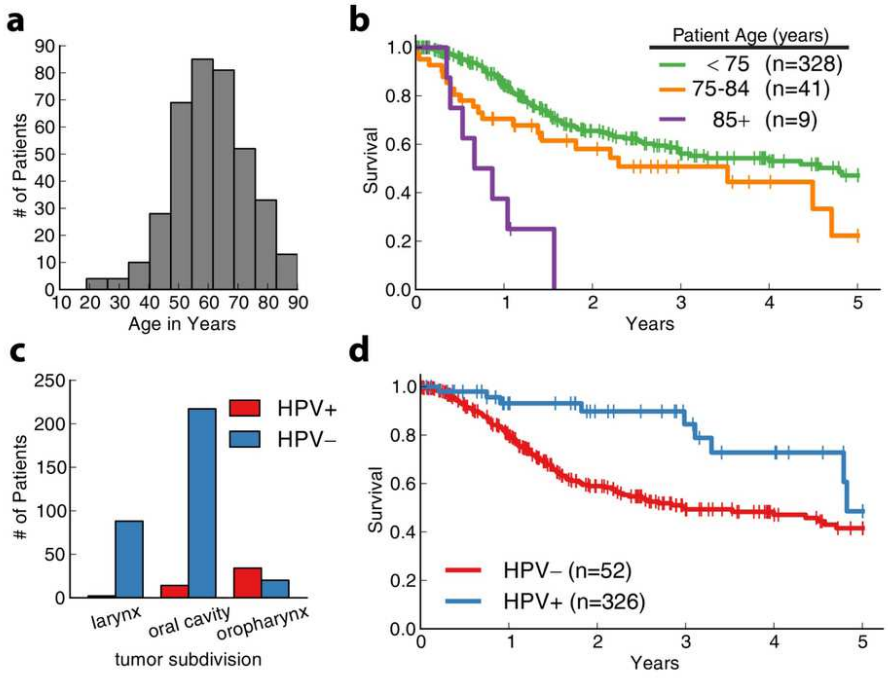
We thank K. Messer and A. Tward for helpful discussions. We gratefully acknowledge support for this study from the National Institutes of Health (P50 GM085764, P41 GM103504 to TI; T32 DC000028 to RO, Burroughs Welcome Fund CAMS to QN; P50 CA097190 and The American Cancer Society to JG; K07CA137140 to AME. J.P.S. is supported in part by grants from the Marsha Rivkin Center for Ovarian Cancer Research and a Conquer Cancer Foundation of ASCO Young Investigator Award.

Chapter 1, in full, is a reprint of material as it appears in *Nature Genetics* 2014. Andrew M Gross, Ryan K Orosco, John P Shen, Ann Marie Egloff, Hannah Carter, Matan Hofree, Michel Choueiri, Charles S Coffey, Scott M Lippman, D Neil Hayes, Ezra E Cohen, Jennifer R Grandis, Quyen T Nguyen, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

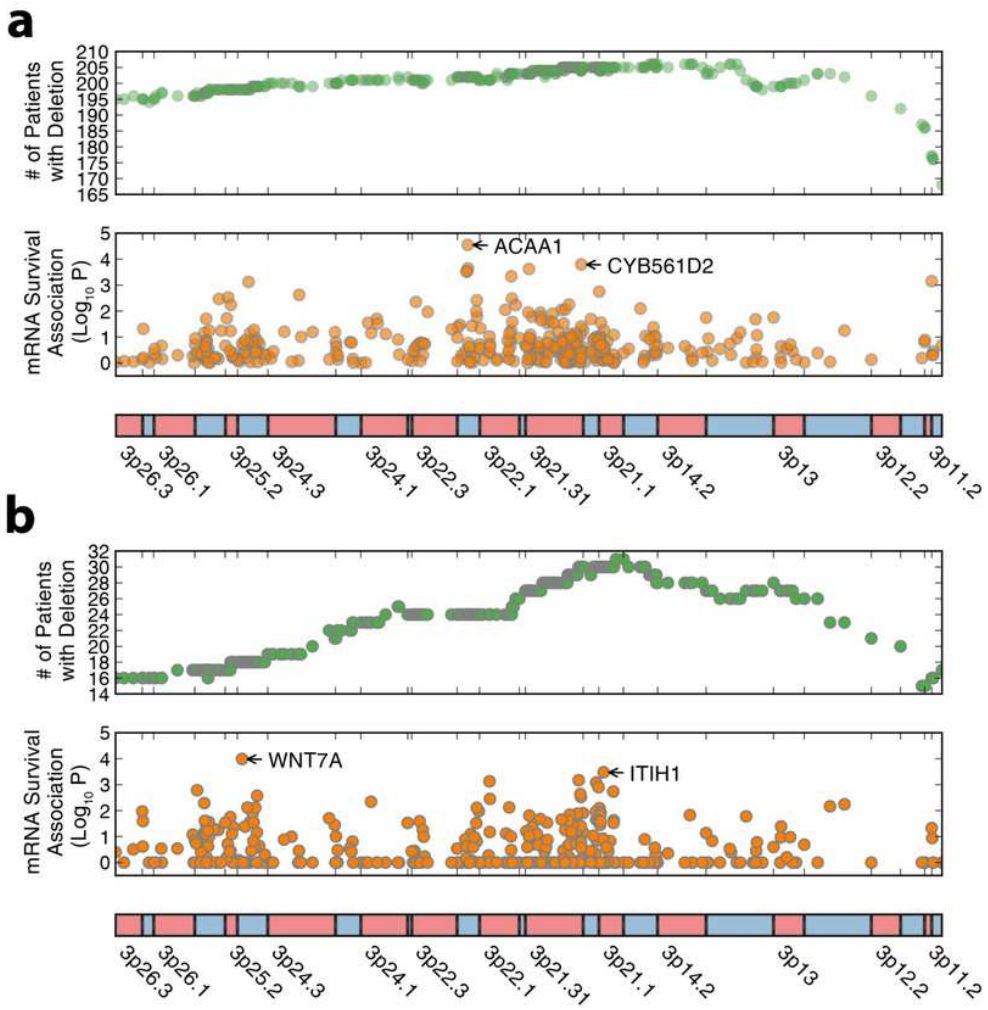
Chapter 1.8: Supplementary Figures



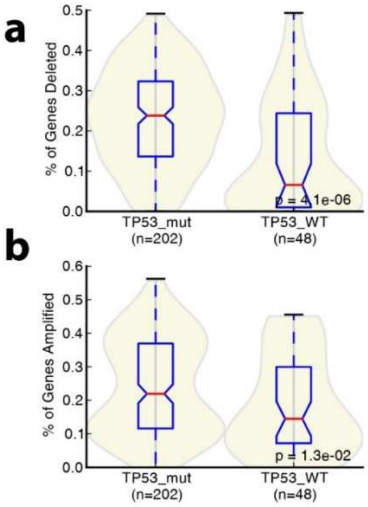
Supplementary Figure 1.1: Integration and selection of cancer events in HNSCC. (a) Tumor data first pass an integration step in which knowledge of pathway or chromosomal structure is used to create meta-features. Data are then filtered based on event frequency across tumor samples or comparison with matched normal samples, yielding a pool of candidate cancer associated events. (b) Example of integration step in which sparse mutations to the SOS1/RAS pathway (Reactome 524) are combined to derive c, a single pathway mutation marker for each patient. In b, green bars represent that a patient (column) has a mutation in a particular pathway gene (row). (d) Example integration of mRNA expression on a pathway, in which Principal Component Analysis (PCA) is applied to the gene-by-patient expression matrix. Shown are the gene loadings for PCA of the PIP3 signaling pathway (mSigDB M1315). (e) In each patient, the first principal component is used to represent the consensus expression value of the pathway. Here the blue bars represent patients for which this value is above threshold, and for which the pathway is scored as 'upregulated'.



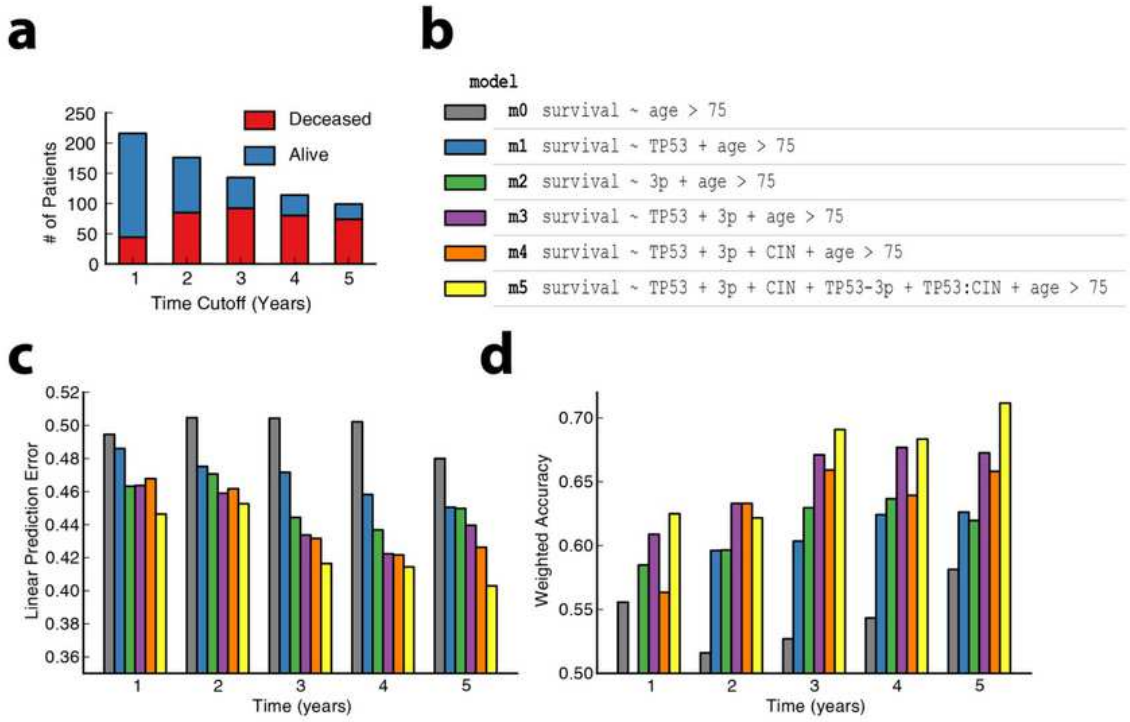
Supplementary Figure 1.2: Characterization of patient age and HPV status in the TCGA HNSCC cohort. (a) Distribution of patient ages across the 378 patients in the cohort. (b) Kaplan-Meyer survival curves for different age cutoffs used in this study. (c) Distribution of HPV+ and HPV+ tumors across different tumor subdivisions. (d) Kaplan-Meyer survival curves comparing HPV+ and HPV+ patients.



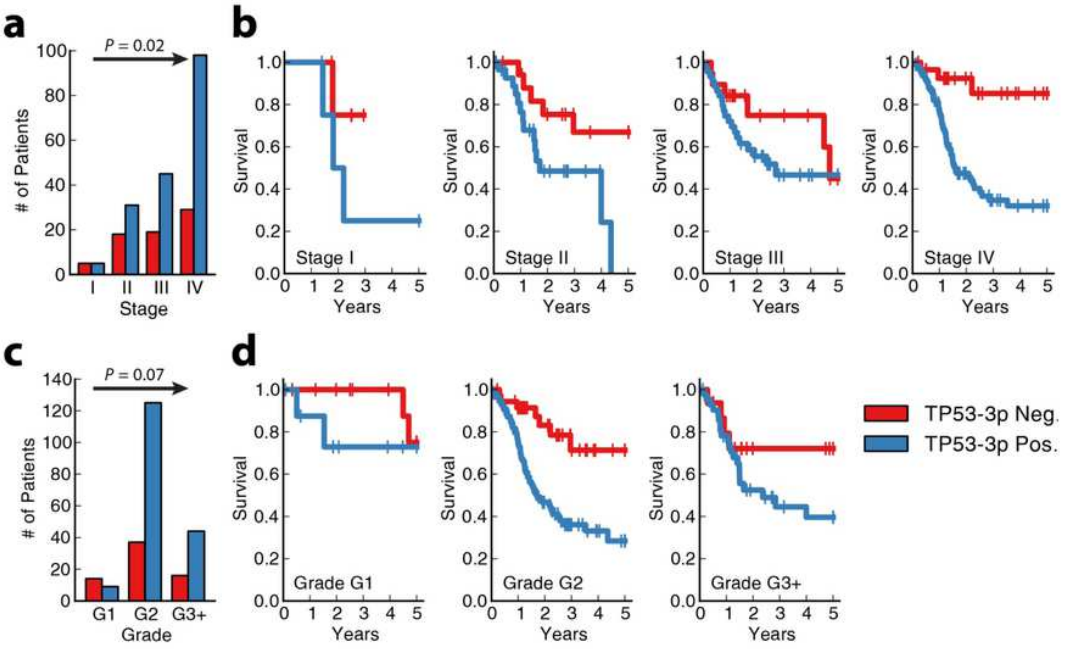
Supplementary Figure 1.3: Exploration of the 3p chromosomal arm. Number of patients with heterozygous loss (top) and association with patient survival (bottom) for genes along the 3p chromosomal arm in TCGA discovery cohort patients with HPV+ (a) and HPV+ (b) tumors.



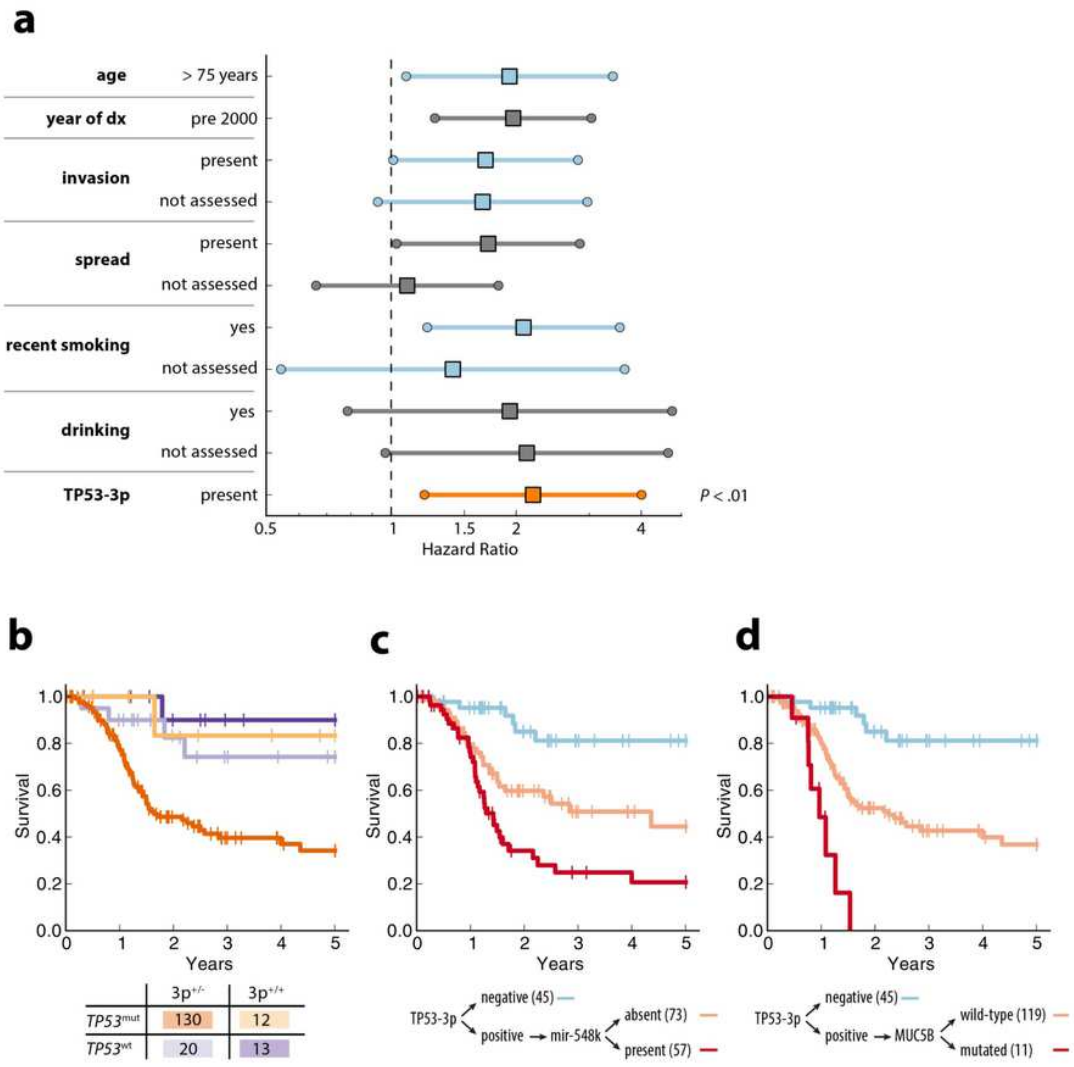
Supplementary Figure 1.4: Exploration of TP53 mutation in the context of chromosomal instability. Violin plots showing the effect of *TP53* mutation on deletion (a) and amplification rates (b). P-values indicate significance of Kruskal-Wallis test.



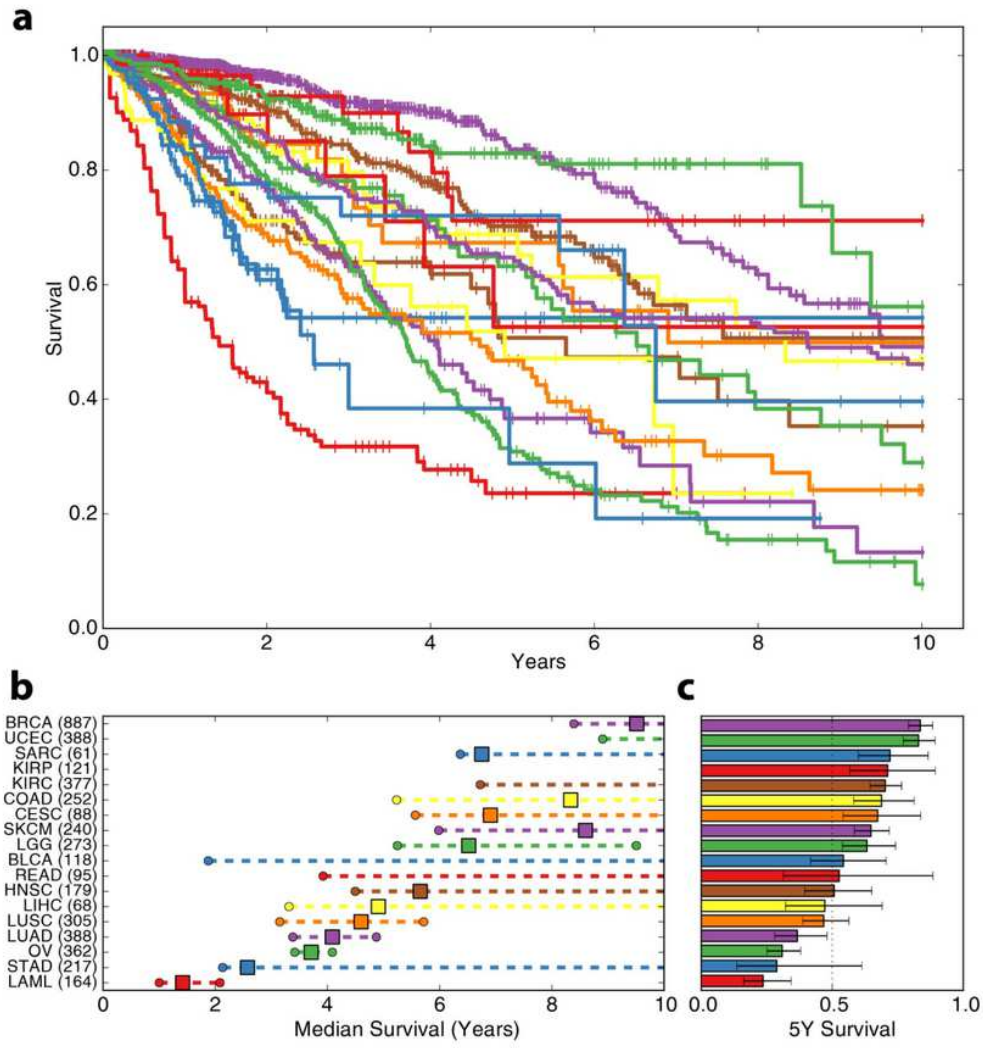
Supplementary Figure 1.5: Exploration of TP53-3p interaction with respect to patient survival. (a) Number of patients surviving or deceased for various time intervals. (b) Statistical models fit using logistic regression. CIN indicates chromosomal instability, measured by the fraction of deleted genes per tumor genome. (c-d) Performance of each logistic regression model in leave-one-out cross-validation to assess ability of different combinations of genomic variables to predict patient outcomes. For description of regression formulation, see Methods. For multivariate Cox analysis of the best model, m5, using the full censored dataset see Supplementary Table 7.



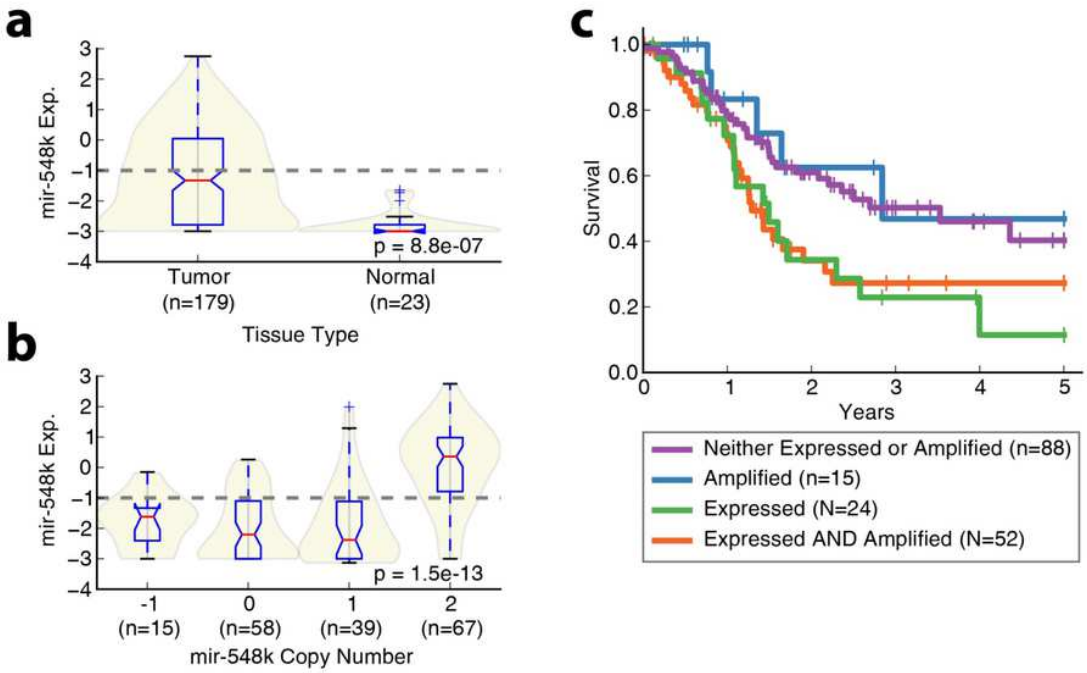
Supplementary Figure 1.6: Subtypes in the context of clinical stage and grade. (a) Frequency and (b) Prognostic effect of *TP53*-3p aggregate event across different stage groups. (c) Frequency and (d) Prognostic effect of *TP53*-3p aggregate event across different grade groups. P-values indicate significance of Kruskal-Wallis test assessing association of *TP53*-3p event with increasing stage or grade.



Supplementary Figure 1.7: Analysis of clinical covariates with molecular subtypes. (a) Hazard Ratios (x-axis) for each component (y-axis) of a multivariate Cox model of patient survival, including *TP53*-3p event and clinical variables. All hazard ratios are relative to absence of the clinical or molecular event. Stepwise feature selection was performed to reduce the model to informative clinical variables only. See Supplementary Table 1 for more information on clinical variables. (b-d) Re-creation of main prognostic associations from this study in a clinically homogenous cohort of 175 patients with a history of smoking and under 75 years of age.



Supplementary Figure 1.8: Pan-cancer analysis. (a) Kaplan-Meyer survival plots, b, median survival, and c, five-year survival for TCGA cancers (error bars indicate 95% CI). Cancer acronyms are defined as follows: BRCA: Breast invasive carcinoma, UCEC: Uterine Corpus Endometrioid Carcinoma, KIRP: Kidney renal papillary cell carcinoma, CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma, LGG: Brain Lower Grade Glioma, COAD: Colon adenocarcinoma, KIRC: Kidney renal clear cell carcinoma, SKCM: Skin Cutaneous Melanoma, SARC: Sarcoma, READ: Rectum adenocarcinoma, LUSC: Lung squamous cell carcinoma, HNSC: Head and Neck squamous cell carcinoma, BLCA: Bladder Urothelial Carcinoma, LIHC: Liver hepatocellular carcinoma, LUAD: Lung adenocarcinoma, STAD: Stomach adenocarcinoma, OV: Ovarian serous cystadenocarcinoma, LAML: Acute Myeloid Leukemia.



Supplementary Figure 1.9: Characterization of mir-548k in patients with the TP53-3p event. (a) Mir-548k expression level in tumor and normal tissues. (b) Comparison of mir-548k copy number with expression. (c) Kaplan-Meyer survival curves of different combinations of high/low mir-548k expression and amplification of its chromosomal segment.

Chapter 1.9: References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300.

Bredel, M. (2009). A network model of a cooperative genetic landscape in brain tumors. J. Am. Med. Assoc. 302, 261–275.

Chung, C.H. (2004). Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell 5, 489–500.

Cibulskis, K. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 22, 398–406.

Cox, D.R. (1984). Analysis of Survival Data.

Farshadpour, F. (2011). Survival analysis of head and neck squamous cell carcinoma: influence of smoking and drinking. Head Neck 33, 817–823.

Lawrence, M.S. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495–501.

Lawrence, M.S., Sougnez, C., Lichtenstein, L., Cibulskis, K., Lander, E., Gabriel, S.B., Getz, G., Ally, A., Balasundaram, M., Birol, I., Bowlby, R., Brooks, D., Butterfield, Y.S.N., Carlsen, R., Cheng, D., Chu, A., Dhalla, N., Guin, R., Holt, R.A., Jones, S.J.M., Lee, D., Li, H.I., Marra, M.A., Mayo, M., Moore, R.A., Mungall, A.J., Gordon Robertson, A., Schein, J.E., Sipahimalani, P., Tam, A., Thiessen, N., Wong, T., Protopopov, A., Santoso, N., Lee, S., Parfenov, M., Zhang, J., Mahadeshwar, H.S., Tang, J., Ren, X., Seth, S., Haseley, P., Zeng, D., Yang, L., Xu, A.W., Song, X., Pantazi, A., Bristow, C.A., Hadjipanayis, A., Seidman, J., Chin, L., Park, P.J., Kucherlapati, R., Akbani, R., Casasent, T., Liu, W., Lu, Y., Mills, G., Motter, T., Weinstein, J., Diao, L., Wang, J., Hong Fan, Y., Liu, J., Wang, K., Todd Auman, J., Balu, S., Bodenheimer, T., Buda, E., Neil Hayes, D., Hoadley, K.A., Hoyle, A.P., Jefferys, S.R., Jones, C.D., Kimes, P.K., Liu, Y., Marron, J.S., Meng, S., Mieczkowski, P.A., Mose, L.E., Parker, J.S., Perou, C.M., Prins, J.F., Roach, J., Shi, Y., Simons, J.V., Singh, D., Soloway, M.G., Tan, D., Veluvolu, U., Walter, V., Waring, S., Wilkerson, M.D., Wu, J., Zhao, N., Cherniack, A.D., Hammerman, P.S., Tward, A.D., Sekhar Pedamallu, C., Saksena, G., Jung, J., Ojesina, A.I., Carter, S.L., Zack, T.I., Schumacher, S.E., Beroukhim, R., Freeman, S.S., Meyerson, M., Cho, J., Chin, L., Getz, G., Noble, M.S., DiCara, D., Zhang, H., Heiman, D.I., Gehlenborg, N., Voet, D., Lin, P., Frazer, S., Stojanov, P., Liu, Y., Zou, L., Kim, J., Sougnez, C., Gabriel, S.B., Lawrence, M.S., Muzny, D., Doddapaneni, H., Kovar, C., Reid, J., Morton, D., Han, Y., Hale, W., Chao, H., Chang, K., Drummond, J.A., Gibbs, R.A., Kakkar, N., Wheeler, D., Xi, L., Ciriello, G., Ladanyi, M., Lee, W., Ramirez, R., Sander, C., Shen, R., Sinha, R., Weinhold, N., Taylor, B.S., Arman Aksoy, B., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Reva, B., Schultz, N., Onur Sumer, S., Sun, Y., Chan, T.A., Morris, L.G., Stuart, J., Benz, S., Ng, S., Benz, C., Yau, C., Baylin, S.B., Cope, L., Danilova, L., Herman, J.G., Bootwalla, M., Maglinte, D.T., Laird, P.W., Triche, T., Weisenberger, D.J., Van Den Berg, D.J., Agrawal, N., Bishop, J., Boutros, P.C., Bruce, J.P., Averett Byers, L., Califano, J., Carey, T.E., Chen, Z., Cheng, H., Chiosea, S.I., Cohen, E., Diergaarde, B., Marie Egloff, A., El-Naggar, A.K., Ferris, R.L., Frederick, M.J., Grandis, J.R., Guo, Y., Haddad, R.I., Hammerman, P.S., Harris, T., Neil Hayes, D., Hui, A.B.Y., Jack Lee, J., Lippman, S.M., Liu, F.-F., McHugh, J.B., Myers, J., Kwok Shing Ng, P., Perez-Ordonez, B., Pickering, C.R., Prystowsky, M., Romkes, M., Saleh, A.D., Sartor, M.A., Seethala, R., Seiwert, T.Y., Si, H., Tward, A.D., Van Waes, C., Waggott, D.M., Wiznerowicz, M., Yarbrough, W.G., Zhang, J., Zuo, Z., Burnett, K., Crain, D., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Sherman, M., Yena, P., Black, A.D., Bowen,

J., Frick, J., Gastier-Foster, J.M., Harper, H.A., Leraas, K., Lichtenberg, T.M., Ramirez, N.C., Wise, L., Zmuda, E., Baboud, J., Jensen, M.A., Kahn, A.B., Pihl, T.D., Pot, D.A., Srinivasan, D., Walton, J.S., Wan, Y., Burton, R.A., Davidsen, T., Demchok, J.A., Eley, G., Ferguson, M.L., Mills Shaw, K.R., Ozenberger, B.A., Sheth, M., Sofia, H.J., Tarnuzzer, R., Wang, Z., Yang, L., Claude Zenklusen, J., Saller, C., Tarvin, K., Chen, C., Bollag, R., Weinberger, P., Golusiński, W., Golusiński, P., Ibbs, M., Korski, K., Mackiewicz, A., Suchorska, W., Szybiak, B., Wiznerowicz, M., Burnett, K., Curley, E., Gardner, J., Mallery, D., Penny, R., Shelton, T., Yena, P., Beard, C., Mitchell, C., Sandusky, G., Agrawal, N., Ahn, J., Bishop, J., Califano, J., Khan, Z., Bruce, J.P., Hui, A.B.Y., Irish, J., Liu, F.-F., Perez-Ordonez, B., Waldron, J., Boutros, P.C., Waggott, D.M., Myers, J., William, W.N., Lippman, S.M., Egea, S., Gomez-Fernandez, C., Herbert, L., Bradford, C.R., Carey, T.E., Chepeha, D.B., Haddad, A.S., Jones, T.R., Komarck, C.M., Malakh, M., McHugh, J.B., Moyer, J.S., Nguyen, A., Peterson, L.A., Prince, M.E., Rozek, L.S., Sartor, M.A., Taylor, E.G., Walline, H.M., Wolf, G.T., Boice, L., Chera, B.S., Funkhouser, W.K., Gulley, M.L., Hackman, T.G., Neil Hayes, D., Hayward, M.C., Huang, M., Kimryn Rathmell, W., Salazar, A.H., Shockley, W.W., Shores, C.G., Thorne, L., Weissler, M.C., Wrenn, S., Zanation, A.M., Chiosea, S.I., Diergaarde, B., Marie Egloff, A., Ferris, R.L., Romkes, M., Seethala, R., Brown, B.T., Guo, Y., Pham, M., and Yarbrough, W.G. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517, 576–582.

Leemans, C.R., Braakhuis, B.J.M., and Brakenhoff, R.H. (2011). The molecular biology of head and neck cancer. Nat. Rev. Cancer 11, 9–22.

Liberzon, A. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740.

Lui, V.W.Y. (2013). Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. Cancer Discov. 3, 761–769.

Marur, S., D'Souza, G., Westra, W.H., and Forastiere, A.A. (2010). HPV-associated head and neck cancer: a virus-related cancer epidemic. Lancet Oncol. 11, 781–789.

McKenna, A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

Mell, L.K. (2010). Predictors of competing mortality in advanced head and neck cancer. J. Clin. Oncol. 28, 15–20.

Meredith, S.D. (1995). Chromosome 11q13 amplification in head and neck squamous cell carcinoma. Association with poor prognosis. Arch. Otolaryngol. Head Neck Surg. 121, 790–794.

Mermel, C.H. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12, R41.

Mroz, E.A. (2013). High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma: Genetic Heterogeneity and HNSCC Outcome. Cancer 119, 3034–3042.

Ohta, M. (1996). The FHIT gene, spanning the chromosome 3p14.2 fragile site and renal carcinoma-associated t(3;8) breakpoint, is abnormal in digestive tract cancers. Cell 84, 587–597.

Partridge, M. (1999). The prognostic significance of allelic imbalance at key chromosomal loci in oral cancer. Br. J. Cancer 79, 1821–1827.

Partridge, M., Emilion, G., and Langdon, J.D. (1996). LOH at 3p correlates with a poor survival in oral squamous cell carcinoma. Br. J. Cancer 73, 366–371.

Perez, F., and Granger, B.E. (2007). IPython: a system for interactive scientific computing. Comput. Sci. Eng. 9, 21–29.

Pickering, C.R. (2013). Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. Cancer Discov. 3, 770–781.

Poeta, M.L. (2007). TP53 mutations and survival in squamous-cell carcinoma of the head and neck. N. Engl. J. Med. 357, 2552–2561.

Rosin, M.P. (2000). Use of allelic loss to predict malignant risk for low-grade oral epithelial dysplasia. Clin. Cancer Res. 6, 357–362.

Song, Y. (2014). Identification of genomic alterations in oesophageal squamous cell cancer. Nature 509, 91–95.

Stransky, N. (2011). The mutational landscape of head and neck squamous cell carcinoma. Science 333, 1157–1160.

Temam, S. (2007). Epidermal growth factor receptor copy number alterations correlate with poor clinical outcome in patients with head and neck squamous cancer. J. Clin. Oncol. 25, 2164–2170.

Thomas, M., Pim, D., and Banks, L. (1999). The role of the E6-p53 interaction in the molecular pathogenesis of HPV. Oncogene 18, 7690–7700.

Walter, V. (2013). Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. PLoS ONE 8, e56823.

Xing, F. (2012). Concurrent loss of the PTEN and RB1 tumor suppressors attenuates RAF dependence in melanomas harboring V600EBRAF. Oncogene 31, 446–457.

(2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120.

Chapter 2: Analysis of Matched Tumor and Normal Profiles Reveals Common Transcriptional and Epigenetic Signals Shared across Cancer Types

Chapter 2.1: Abstract

To identify the transcriptional regulatory changes that are most widespread in solid tumors, we performed a pan-cancer analysis using over 600 pairs of tumors and adjacent normal tissues profiled in The Cancer Genome Atlas (TCGA). Frequency of upregulation was calculated across mRNA expression levels, microRNA expression levels and CpG methylation sites and is provided here as a resource. Frequent tumor-associated alterations were identified using a simple statistical approach. Many of the identified changes were consistent with the increased rate of cell division in cancer, such as the overexpression of cell cycle genes and hypermethylation of PRC2 binding sites. However, we also identified proliferation-independent alterations, which highlight novel pathways essential to tumor formation. Nearly all of the GABA receptors are frequently downregulated, with the gene encoding the delta subunit (GABRD) strongly upregulated as the notable exception. Metabolic genes are also frequently downregulated, particularly alcohol dehydrogenases and others consistent with the decreased role of oxidative phosphorylation in cancerous cells. Alterations in the composition of GABA receptors and metabolism may play a key role in the differentiation of cancer cells, independent of proliferation.

Chapter 2.2: Introduction

Cancerous cells are characterized by numerous changes to the genome, epigenome, transcriptome. While most tumor-associated changes have little function, key genes and pathways are often implicated by looking across patients within a cohort for events that are recurrent (Ding *et al.*, 2008; McLendon *et al.*, 2008; Lawrence *et al.*, 2013). While such analyses are traditionally performed across well-defined patient populations with tumors of similar anatomical location and histological appearance, large data sets produced by public efforts such as The Cancer Genome Atlas (TCGA) (McLendon *et al.*, 2008; Chang *et al.*, 2013) have now made meta-analysis of cancer studies feasible.

By looking across many different subtypes, pan-cancer analyses provide a high level, tissue agnostic view of cancer. Many such studies have analyzed coordinated changes across molecular phenotypes and clinical data to isolate key signals during tumorgenesis. Such efforts have uncovered conserved patterns of gene co-expression across many types of tumors (Segal *et al.*, 2004; Cheng *et al.*, 2013) identifying molecular patterns associated with tumor growth and proliferation. In a complementary approach, a recent paper by Gentles and colleagues (Gentles *et al.*, 2015) identified genes whose expression was associated with survival across cohorts spanning many tissues. These authors found that the overexpression of genes near the FOXM1 transcriptional network and of genes that drive cell cycle progression were associated with adverse patient outcomes. These highly conserved signatures of cell proliferation support the

hypothesis that a core cancer phenotype is activated to varying degrees across diverse tumor types.

Thus far, such pan-cancer studies of transcriptional changes have focused mainly on tumor samples, without consideration of normal tissue. In contrast, studies of mutations, structural variations or DNA copy number alterations have frequently relied on subtractive analysis of matched data to achieve power in detecting tumor-specific changes. Although a few expression studies analyzed patient-matched tumors and adjacent normal tissue, these studies were restricted to specific tissue cohorts (Gardina *et al.*, 2006; Hamfjord *et al.*, 2012; Seo *et al.*, 2012; Notterman *et al.*, 2001; Kobayashi *et al.*, 2011; Terunuma *et al.*, 2014). They were thus capable of identifying genes whose expression in tumor deviates from normal in a single tissue, but were unable to distinguish which of these changes are specific to a given study population or are general features of cancer as a whole. To this effect, a pan-cancer analysis of differential transcriptional regulatory programs—whether at the level of mRNA expression, miRNA expression or methylation—has not yet been performed.

Here, we perform such an analysis using information readily available in The Cancer Genome Atlas (TCGA), which has enabled standard data collection procedures and molecular profiling assays for numerous measurement platforms (Chang *et al*, 2013). Using TCGA data, we compile a comprehensive list of tumor-associated mRNAs, miRNAs and methylation sites by measuring the frequency at which their levels are elevated between matched tumor and normal samples across all measured cancer tissues. The upregulation frequencies for these

features are provided as a general resource to the cancer community. We find that in addition to near-universal overexpression of genes important for tumor proliferation, there exist prominent proliferation-independent signals which could play a role in tissue remodeling.

Chapter 2.3: Results

To identify ubiquitous tumor-associated signals, we downloaded all of the available data from TCGA as of April 2, 2015, through the Broad Institute's Firehose web portal (Methods) (Broad Institute Genome Data Analysis Center, 2015). This dataset consisted of genome-wide mRNA expression, microRNA (miRNA) expression and CpG methylation for over 9,000 tumors, of which adjacent normal tissues were also profiled for over 600 patients (**Supplementary Figure 2.1**).

Given this large collection of matched tumor and normal data, we were powered to employ a simplified analysis to identify molecular signals associated with tumors (**Methods**, **Figure 2.1a** and **Supplementary Figure 2.2**). For each mRNA, miRNA or CpG marker, we quantified fraction upregulated (f_{up}), the fraction of patients for which the marker level was higher in the tumor than in the matched normal tissue. This metric is a formulation of the sign-test statistic $p = \Pr(\mathbf{x}_i > \mathbf{y}_i)$, where **x** and **y** are vectors of matched samples from tumor and adjacent normal tissue, respectively. Using this statistic we identified mRNAs, miRNAs and CpGs that ranged from random ($f_{up} = 0.5$) to highly differentially expressed or methylated (f_{up} approaching 0 or 1) (**Figure 2.1b** and **Supplementary Table 2.1**). To assess

the reproducibility of this statistic, we studied 10 additional gene expression microarray datasets, spanning 1012 subjects with matched tumor/normal data from the Gene Expression Omnibus. After calculating f_{up} for all of the genes in the dataset, we found a correlation of 0.84 ($P < 10^{-16}$, 95% confidence interval (CI): 0.838–0.847) between these scores and the f_{up} scores identified from TCGA RNA-sequencing data (**Figure 2.1c** and **Supplementary Table 2.2**).

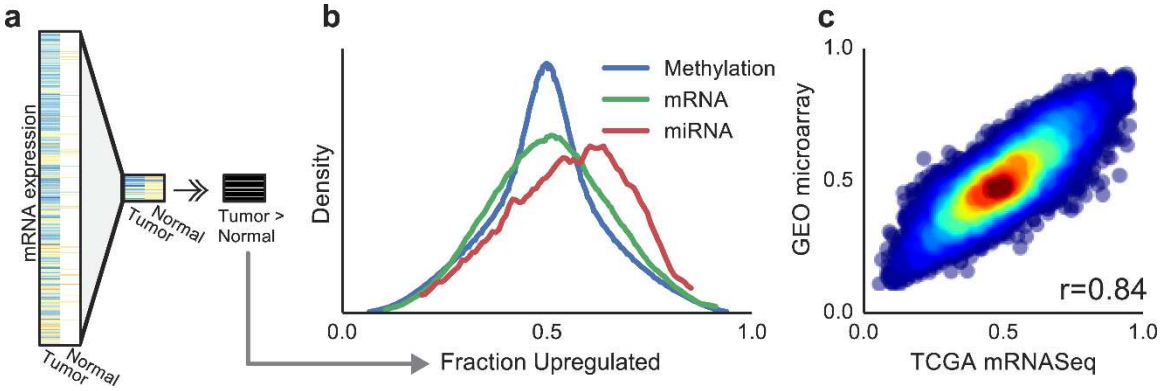


Figure 2.1: Description of the fup statistic. (a) Schematic of the calculation of fraction upregulated (f_{up}) for a single gene expression profile across the TCGA cohort. Data are filtered to include only matched samples, the magnitudes of paired tumor/normal samples are compared, and a fraction of how often the gene is upregulated is recorded. (b) Density of f_{up} statistic across genome-wide mRNA, miRNA, and methylation measurements. (c) Comparison of mRNA f_{up} statistic calculated from TCGA mRNASeq measurements versus microarray measurements downloaded from GEO.

Inspection of molecular entities with extreme values of f_{up} confirmed that tumor proliferation plays a dominant role, as described by previous studies (Segal *et al.*, 2004; Cheng *et al.*, 2013; Gentles *et al.*, 2015; Evan and Vousden, 2001; Wierstra and Alves, 2007). Among the most heavily tumor-associated genes was *FOXM1*, for which the mRNA levels are upregulated in 93% of patient tumors (95% CI_{Bonf}: 87%-97%). *FOXM1* is a well-known proliferation-associated transcription

factor which plays a central role in regulating the progression of the cell cycle (Wierstra and Alves, 2007). Gene-Set Enrichment Analysis highlighted a number of features associated with proliferation, including upregulation of cell cycle genes with particularly large effect sizes observed for the cell cycle gene subsets "deposition of CENPA containing nucleosomes at the centromere" and "M/G1 transition" (**Figure 2.2a** and **Supplementary Table 2.3**, Mann-Whitney U test, $P_{BH} < 10^{-16}$). Analysis of methylation markers showed hypermethylation occurring at PRC2 binding sites which have been previously linked to proliferation in cancer (Margueron and Reinberg, 2011) (**Figure 2.2b**). Taken together, these findings confirm that many tumor-associated molecular changes are driven by proliferation.

To isolate proliferation dependent and independent components of the tumor associated signal, we assigned a proliferation score for each mRNA, miRNA and methylation site. This was calculated by assessing the correlation across TCGA patients of each feature expression level with a previously published proliferation signature (Venet *et al.*, 2011) (meta-PCNA, Methods). Indeed we found that these proliferation scores were highly correlated with fup scores across all three data types, with Pearson's $r = 0.63$ (95% CI: 0.62–0.64), 0.62 (0.56–0.67), and 0.674 (0.672–0.676) for mRNA, miRNA and methylation, respectively (**Figure 2.2c**, for all three statistics $P < 10^{-16}$). Interestingly, we observed a heavy skew in the fup statistic for miRNA species in particular (**Figure 2.1a**), which we attribute to a general trend of increasing miRNA expression with proliferation (Bueno *et al.*, 2008).

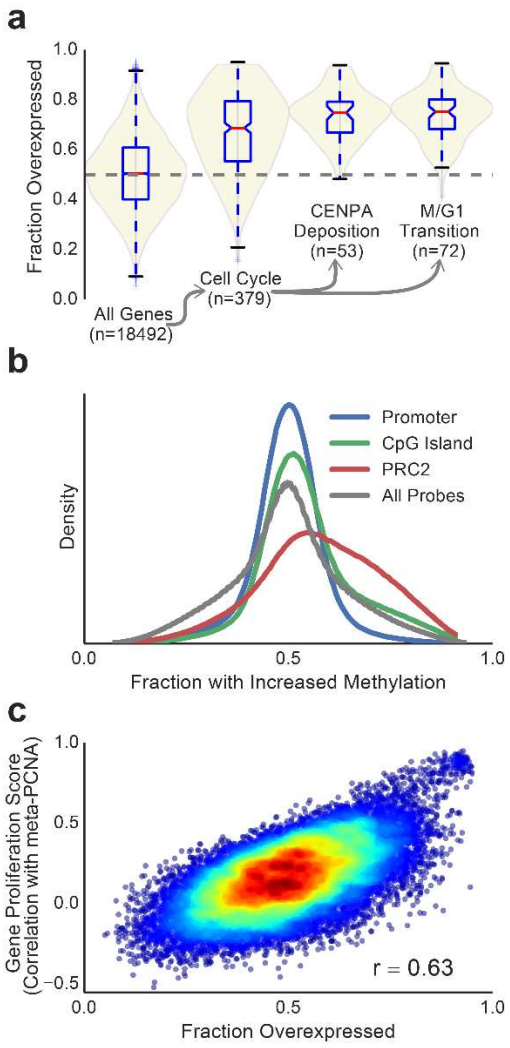


Figure 2.2: Tumor-associated features are consistent with proliferative signals. (a) Violin plots showing distribution of mRNA level f_{up} statistic (fraction overexpressed) across all genes, compared to genes annotated to the cell cycle and its subsets: "deposition of CENPA containing nucleosomes at the centromere" and "M/G1 transition" in mSigDB. (b) Density plots of the distribution of f_{up} (fraction with increased methylation) across methylation markers annotated to functional genomic sites. (c) Scatter plot comparing f_{up} statistic against gene correlation with proliferation for every gene expression profile.

To assess tumor-associated, growth-independent signals, we adjusted marker levels to remove any association with proliferation and recalculated fup (i.e., accounting for the meta-PCNA signature, see **Methods**, **Supplementary Table 2.4**). We expected that features with extreme values of detrended fup would

be altered in the transition from normal to tumor cells, but not associated the tumor growth rate. Enrichment analysis of this detrended statistic identified overexpression of genes involved in ribosomal and proteasomal processes (**Supplementary Table 2.5**, Mann-Whitney U test, $P_{BH} < 10^{-16}$, $P_{BH} < 10^{-7}$, respectively). Interestingly, while telomere maintenance genes had a general increase in fup, genes involved with telomere extension had much stronger correlations with proliferation than genes involved in packaging of telomere ends ($P < 0.001$, **Supplementary Figure 2.3**). It is likely that these and other pathways are important for the initial rewiring of the cell required for accelerated growth but then have little impact on the tumor's growth rate.

The most upregulated, proliferation-independent genes in tumors were *SEMA5B* (detrended f_{up} = 0.82 [0.74–0.88], **Supplementary Figure 2.4**), the GABA receptor subunit *GABRD* (detrended f_{up} = 0.82 [0.64–0.80], **Figure 2.3**), and the well-studied tumor suppressor CDKN2A (detrended f_{up} = 0.72 [0.63–0.79]). *SEMA5B* is a gene in the semaphorin family, whose main roles are to serve as guidance signals in various stages of development. These genes have recently been shown have a role in cancer signaling (Tamagnome, 2012). This GABAA subunit is primarily expressed in the cerebellum where its receptor is located extrasynaptically (Nusser *et al.*, 1995; Mele *et al.*, 2015), but it is also expressed in the testes (**Supplementary Figure 2.5**) and CD4+ T-cells (Mele *et al.*, 2015, Tian *et al.*, 2004). In the TCGA dataset, GABRD is overexpressed in 89% (CI_{Bonf} 81%-93%) of subjects and has a slight negative association with proliferation in tumors (**Figure 2.3**). In contrast, most other GABA subunit genes are

downregulated across many cancers (**Figure 2.3c**, **Supplementary Figure 2.6**). We observed a particularly large effect in renal cell carcinoma where there is a ten-fold median decrease in *GABRA2* alongside a six-fold increase in expression of *GABRD* (**Figure 2.4e**). Similar effects were observed in a paired microarray dataset (**Supplementary Figure 2.7)**.

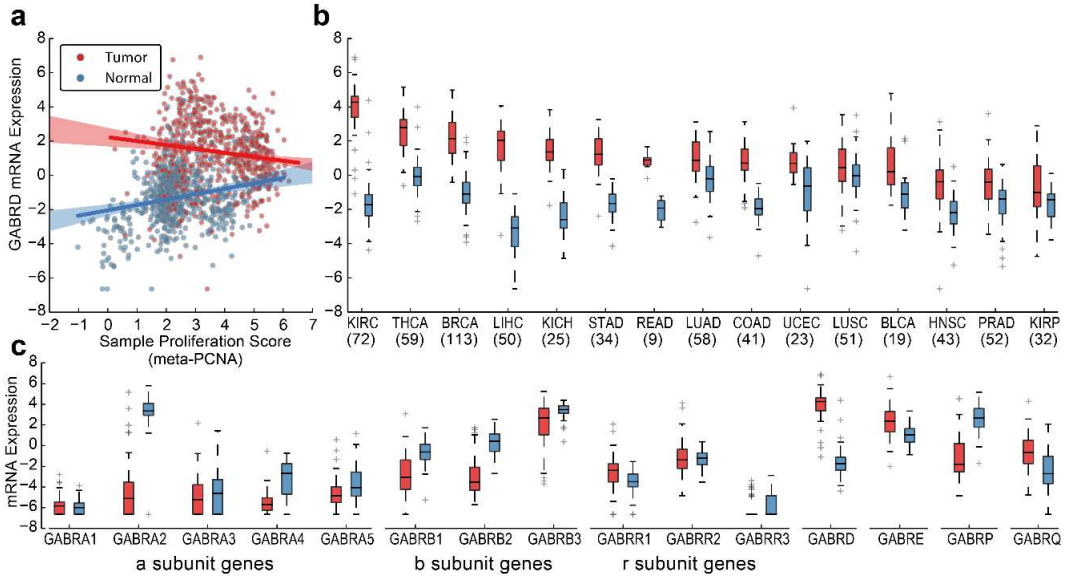


Figure 2.3: GABRD is tumor-associated, independent of proliferation. (a) Scatter-plot comparing GABRD gene expression profiles to proliferation scores across matched tumor and normal samples. Lines indicate linear regression figs of tumor (red) and normal (blue) samples, shaded regions indicate 95% confidence intervals. (b) Comparison of matched tumor and normal profiles for GABRD expression, grouped by tissue type. (c) Comparison of matched tumor and normal profiles for all GABA protein subunits in renal cell carcinoma. Cancer acronyms are defined as follows: KIRC, kidney renal clear cell carcinoma; THCA, thyroid carcinoma; BRCA, breast invasive carcinoma; LIHC, liver hepatocellular carcinoma; KICH, kidney chromophobe; STAD, stomach adenocarcinoma; READ, rectum adenocarcinoma; LUAD, lung adenocarcinoma; COAD, colon adenocarcinoma; UCEC, uterine corpus endometrioid carcinoma; LUSC, lung squamous cell carcinoma; BLCA, bladder urothelial carcinoma; HNSC, head and neck squamous cell carcinoma; PRAD, prostate adenocarcinoma; KIRP, kidney renal papillary cell carcinoma.

Gene sets with similar patterns of differential expression as *GABRD* included ‘hematopoietic cell lineage’ and ‘helper T-cell polarization’ (**Methods**).

Further inspection of genes in the helper T-cell polarization pathway showed a preference for genes expressed in Th1 as opposed to Th2 cells. To determine whether this signal represented infiltration by immune cells into the tumor, we used the CIBERSORT program (Gentles *et al.*, 2015) to predict immune cell subsets in tumor samples, but found little to no association with *GABRD*. While it remains difficult to completely rule out immune infiltration as a driving force of this signal, these findings suggest that increased levels of the delta subunit could lead to functional changes in the GABAA receptor that may play a role in tumor cell differentiation.

Among the most downregulated, proliferation-independent genes we noticed widespread epigenetic silencing in tumors with strong enrichments for transcription start site hypermethylation (**Methods**, **Supplementary Figure 2.8a**, Odds-Ratio = 2, $P < 10^{-16}$) and gene body hypomethylation (**Supplementary Figure 2.8b**, Odds-ratio = 2.5, $P < 10^{-16}$). While coverage of methylation markers on the Illumina 450k chip varied across genes, manual inspection (**Methods**) of the most consistently downregulated genes identified many genes with associated with methylation changes to their DNA including *GSTM5* (detrended f_{up} = 0.27 [0.19–0.35], **Supplementary Figure 2.8c**) and *NRXN1* (detrended f_{up} = 0.25 [0.18–0.34], **Supplementary Figure 2.8d**). While *NRXN1* is primarily expressed in brain where it serves as a cell surface protein, it has also been shown to play a role in remodeling of vascular tissue indicating it may play a wider role in regulation of cell adhesion in the periphery (Bottos, *et al.*, 2009).

A screen for gene-sets enriched for proliferation-independent downregulation identified transcription and fatty acid metabolism pathways (Mann-Whitney U test, $P_{BH} < 10^{-8}$, $P_{BH} < 10^{-4}$, respectively). Among the fatty acid metabolism gene set were the alcohol dehydrogenase genes which were nearly ubiquitously down-regulated with a particularly large effect for the class I genes (f_{up} = 0.06 [0.02–0.10], 0.05 [0.02–0.10] and 0.12 [0.06–0.18] for *ADH1-A, -B* and -C, respectively) as well as *ALDH2* (f_{up} = 0.15 [0.09–0.22]), which serves to break down acetaldehyde (**Figure 2.4** and **Supplementary Figure 2.9**). The downregulation of alcohol metabolism is likely a component of alternative pyruvate usage mediated by the Warburg effect in which cancer cells increase their rate of glycolysis by shifting to aerobic metabolism (Warburg, 1956). Exploration of other glycolysis genes supported this shift with upregulation of the lactate dehydrogenase gene *LDHA* (f_{up} = 0.79 [0.71–0.86]) alongside downregulation of the mitochondrial pyruvate carrier gene *MPC1* (f_{up} = 0.11 [0.09–0.22], TCGA symbol *BRP44L*). Much like the ADH genes, *MPC1* is downregulated in a proliferation-independent manner, and has recently been shown to affect cancer cell line growth in nonadherent, 3D culture conditions but not in proliferation or cell-cycle progression assays (Schell *et al.*, 2014).

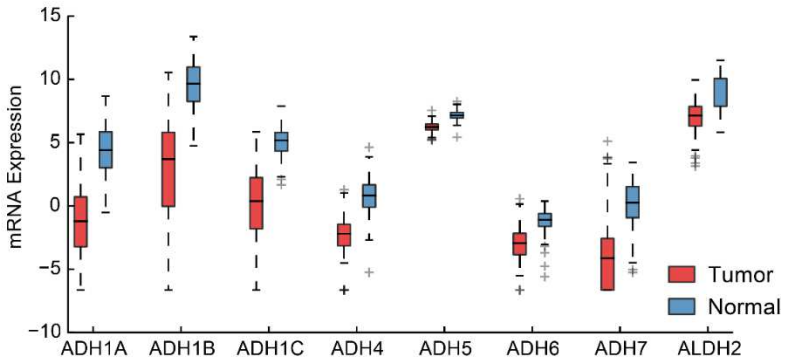


Figure 2.4: Differential expression of alcohol dehydrogenase family of genes. Shown here for TCGA breast cancer dataset as a representative cohort. Also shown is ALDH2 which is the major enzyme responsible for breaking down acetaldehyde, the primary intermediate product of alcohol metabolism.

Chapter 2.4: Discussion

Here we have provided a resource to aid in the understanding of tumor-associated molecular changes. Using the largest database of molecular profiles from paired tumor and adjacent normal tissues available, we determined how often each mRNA, miRNA and methylation site is differentially expressed in cancer.

We observed changes in the expression levels of features associated with growth and proliferation, including cell cycle genes, global miRNA expression and methylation of PRC2 binding sites. In addition to features consistent with rapid cellular proliferation, we also observed a number of proliferation-independent signals. These genes may lie in pathways required for cells to break free of the normal mechanisms that regulate properties such as telomere processing and tissue invasiveness. Such a proliferation-independent pattern could also arise for tumor suppressors. Many tumor suppressors are activated in response to DNA damage but may be actively suppressed by altered molecular signaling in tumors.

One major finding of this study is the proliferation-independent upregulation of *GABRD* in nearly all tumors profiled. In addition to its well-known role of neurological signaling, signaling via GABA subunits can also suppress the proliferation of both neural and peripheral stem cells. In addition, dysregulation of GABA signaling has been implicated in various cancers, where it is hypothesised to have a role in the differentiation and proliferation of tumor stem cells (Young and Bordey, 2009).

There are a number of possible explanations for why many GABA subunits are downregulated, but *GABRD* in particular is upregulated, in cancer. One possibility is that tumors express a novel receptor configuration; another is that the expression of the delta subunit could create non-functional receptors with other subunits. While it is hard to rule out the former explanation, the expression of *GABRD* in the testes (**Supplementary Figure 2.5**), and the observation that GABA has been shown to promote proliferation of Leydig cells in rodent testes (Geigerseder *et al.*, 2003), gives some weight to the idea that usage of an alternative GABAA receptor may be important for tumorigenesis.

Further work is clearly needed to understand the proliferation-independent genes and expand on their role in cancer. While secondary validation methods often measure the change of a cell line's growth rate in response to disruption of a target, phenotypes such as those described here would not likely manifest in such assays. In contrast, non-traditional assays such as cell migration and 3D cell culture may be required to validate such phenotypes. 3D cell culture experiments have recently been conducted on the pyruvate carrier MPC1 in which the

coauthors show a clear induction of growth only when this gene is re-expressed in 3D culture and mouse xenograft models, not in classical (2D) cell culture (Schell *et al.*, 2014).

Finally we would like to highlight the utility of using a large, diverse cohort to derive a robust pan-cancer signal. It is important to note that we do not aim to diminish the importance that normal tissue function, exposure to carcinogens, and cell turnover rates can have on the phenotypes of different cancer presentations. However, signals that are robust to tissue and environmental context are likely to be very important to the core processes driving a broad spectrum of cancer types. With the recent attention towards precision medicine, it is all the more important to define the standard molecular phenotype for cancer in general: Only by first defining common molecular features can we truly understand how treatment can be catered to detect and attack specific presentations of the disease.

Chapter 2.4: Methods

Chapter 2.4.1: Informed Consent

Informed consent was obtained for all patients as part of the Cancer Genome Atlas consortia. All data used in this study were downloaded from public websites after the data were consented for public use. No handling of personally identifiable information was done by the researchers on this study.

Chapter 2.4.2: Molecular Data Retrieval and Processing

All data were downloaded using the Broad Institute's firehose_get data-retrieval utility. To maintain the coherency of the analysis across different data layers and cancer types, we used Level 3 normalized molecular data as the input to our analysis and used all data available as of the April 2, 2015 standard data run. The use of the TCGA Genome Data Analysis Center (GDAC) pipeline is intended to make these results easy to update as more TCGA data become available.

For TCGA gene expression values, we used data provided by Rahman and colleagues, who reprocessed the RNA sequence based expression data and showed better performance on controls (Rahman *et al.*, 2015). While using this data as opposed the standard TCGA pipeline yielded slight changes to the results presented here, they are qualitatively very similar for both pipelines. To maintain consistency and respect data versioning we only used patients and genes present in the Firehose dataset.

A marker (gene, miRNA, methylation probe) filter was applied to TCGA data to ensure that there was a detectable change in value between patient matched tumor and normal profiles in at least 50% of subjects. In general, this approach removed features whose levels were below the limit of detection in both tumor and normal, resulting in identical low values. The resulting feature set consisted of 396,059 methylation probes, 520 microRNA, and 18420 genes.

Microarray data was retrieved via manual search of the Gene Expression Omnibus (GEO) for large molecular cohorts with paired tumor/normal expression

data from the following accessions: GSE25097, GSE14520, GSE62872, GSE44076, GSE53757, GSE39791, GSE5364, GSE41258, GSE39004, GSE68468 and GSE33532. Data were obtained from the pre-processed series matrix files made available on GEO, and probes were averaged onto their annotated genes. Due to the unbalanced distribution of tissues available on GEO, fraction upregulated (f_{up}) statistics were calculated for each tissue type individually, and then averaged to obtain a consensus. As not all microarray platforms had full coverage of the coding genes, statistics were calculated for available data, and genes profiled in fewer than 500 matched samples were discarded. This resulted in 16785 genes for which both microarray and RNA-sequencing data were available.

Chapter 2.4.3: Assessment of Differential Expression via the Fraction of Upregulated Patients

The fraction upregulated metric is a formulation of the sign-test statistic $p = \Pr(\mathbf{x_i} > \mathbf{y}_i)$, where **x** and **y** are vectors of matched samples. This statistic can be seen as a simplification of the Wilcoxon signed rank test, as it does not use the magnitude of the differences for a ranking but rather counts the signs of the differences. This is a simple, assumption-free metric in which information on the magnitude of differential expression or methylation is discarded. The statistic represents the fraction of patients for which a marker takes on a higher value in the tumor than the matched normal sample and ranges between 0 and 1. Statistical assessment of fup is conducted by testing against the null hypothesis that fup assumes a binomial distribution with a mean of 0.5. Confidence intervals are

assessed via examination of a beta distribution fit with shape parameters defined by the sign test. Although such a procedure can greatly limit statistical power when the sample size is small, at large sample sizes, f_{up} tracks very well with parametric statistics such as a paired t-test (**Supplementary Figure 2.2**).

By simplifying to a sign test we lose statistical power, but gain robustness of the test by allowing for application of this test regardless of the distribution of the data. This is used in replacement of standard statistical techniques used such as a paired t-test or specialized differential expression tools which pool variance across markers that are traditionally used in studies that have much smaller sample sizes (generally n = 3–20) and thus lack the power to use such a simplified model. We refrain from using such techniques as they would introduce a wide variety of confounding factors which would make our analysis much less robust and harder for the reader to interpret. For example the use of a t-test without modeling tumor purity as a covariate would be inappropriate in this setting as more pure samples would have an outsized effect.

Furthermore this nonparametric exact test has a number of desirable properties for integrative analysis across datasets. Statistically it relies on no assumptions and is robust to outliers. Furthermore it does not pool samples as biological replicates and thus gives all samples equal weights when calculating a summary value. Biologically the sole assumption of the test is that the tumor sample contains more tumor cells than the normal sample. Due to these properties, we expect little contribution of non-cancer tissue-specific expression and batch effects.

Chapter 2.4.4: Proliferation Scoring

A patient level proliferation score was adopted from the meta-PCNA metric published in Venet *et al.* (2011). This previous study mined normal, non-diseased tissues and defined a set of 131 genes associated with the well-studied Proliferating Cell Nuclear Antigen (*PCNA*) gene, then created a meta-gene calculated as the median expression level of these 131 genes. As in Venet *et al.*, the median of these genes was used to construct the proliferation score in the current study. A marker-level association with this proliferation score was then computed for each gene, miRNA or methylation probe by assessing the Pearson correlation of the change in meta-PCNA with the change in marker levels from tumor to normal tissue for all subjects with matched samples.

Chapter 2.4.5: Assessment of Proliferation-Independent Tumor-Associated Features

To search for features that are tumor-associated independently of proliferation, the association of marker levels with proliferation (meta-PCNA) was detrended via a linear model. The detrended f_{up} metric is very similar to the standard f_{up} calculation with the addition of preprocessing to remove the trends of proliferation. Additional tissue and interaction terms are added to model to association of metaPCNA with tissue.

The detrending step is implemented in R using the following model:

Marker_level ~ metaPCNA + tissue + metaPCNA:tissue

Where metaPCNA:tissue is an interaction term between these two factors. After this model is fit for all markers we obtain a matrix of residuals from the set of

markers, and repeat the screen for conserved changes as previously implemented for f_{up}. The screen result provides us with p-values and confidence intervals for all detrended f_{up} values.

Chapter 2.4.6: Gene Set Enrichment Analysis

Gene sets were downloaded from the Molecular Signatures Database (mSigDB) (Subramanian *et al.*, 2005). Version 5 of the canonical pathway gene sets was used in this analysis. Enrichment of f_{up} for gene sets was performed by screening all sets for a difference in the distribution of f_{up} within the set as compared to the background gene set via the rank-based Mann-Whitney U test.

To understand whether GABRD had coordinated differential expression with any annotated pathways, we conducted an enrichment test against the co-differential expression of GABRD with all other genes. To address this, we assessed enrichment of co-differential expression by the following method:

- dx: gene x gene correlation across matrix of differential expression
- dt: gene x gene correlation across matrix of tumor-only gene expression
- cx: dx-dt, change in correlation
- pathway enrichment: change in mean of cx within genes annotated to a given pathway

During preliminary analysis we noted that proliferation associated pathways were enriched for co-differential expression with many genes. We suspect this is the case due to the strong proliferation component of the differential expression signal giving these genes more information content. To hone in on pathways with a specific enrichment for *GABRD* we computed pathway enrichments for all genes, and ranked *GABRD* with respect to all other genes. For the two pathways

highlighted in the text, ‘hematopoietic cell lineage’ and ‘helper T-cell polarization’ the enrichment of *GABRD* was ranked 3rd and 9th of all genes profiled.

mSigDB pathway IDs for gene sets cited in the main text are as follows:

- cell cycle: M5336
- deposition of CENPAcontaining nucleosomes at the centromere: M871
- M/G1 transition: M10080
- hematopoietic cell lineage: M6856
- helper T-cell polarization: M4047
- ribosome: M189
- proteasome: M10680
- packing of telomere ends: M17695
- telomere extension: M14804
- telomere maintenance: M4052

Chapter 2.4.7: Integration of Methylation and Expression Data-Layers

To understand epigenetic silencing of frequently downregulated genes, we integrated data from the DNA methylation and gene expression data-layers. This analysis took place on the 357 patients with both data-types profiled across tumor and normal tissue samples. Genes were annotated as up- or down-regulated by the significance of the detrended f_{up} metric with a threshold of $P_{Bonf} < 0.05$. The odds-ratio statistic in the main text was constructed by comparing the frequency at which methylation probes were greater or less than the median value of the distribution for probes mapping to downregulated genes against all other probes. To further explore epigenetic silencing, we manually inspected the 10 most proliferation-independent downregulated genes. While multiple-hypothesis corrected p-values of associations are not reported in **Supplementary Figure 2.8**,

we estimate test space to be on the order of 100 tests as 10 genes were explored and around 10 possible combinations of annotations could be constructed.

Chapter 2.4.8: Availability

All data retrieval and processing steps are documented in a series of IPython notebooks (Perez and Granger, 2007) available online (https://github.com/theandygross/TCGA_differential_expression). These notebooks provide fully executable instructions for the reproduction of the analyses and the generation of figures and statistics for this study.

Chapter 2.5: Author Contributions

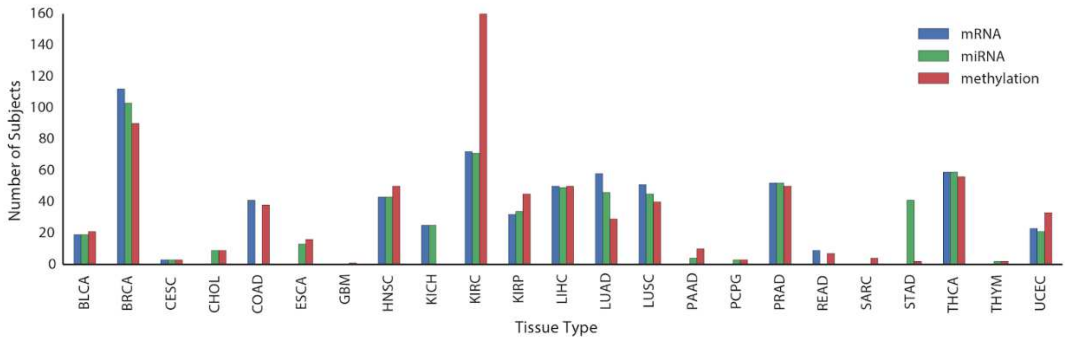
Conceived and designed the experiments: AMG TI. Analyzed the data: AMG. Wrote the paper: AMG JFK TI.

Chapter 2.6: Acknowledgements

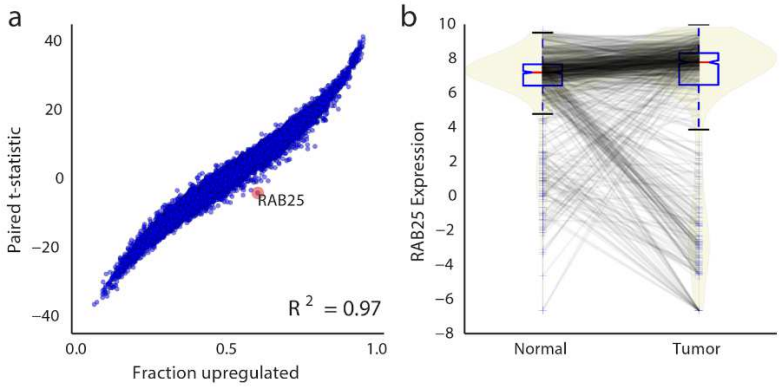
The results published here are based upon data generated by the TCGA Research Network. We would like to thank Hannah Carter and John Paul Shen for helpful discussion and review of the manuscript.

Chapter 2, in full, is a reprint of material as it appears in *PLOS One*, 2015. Andrew M. Gross, Jason F. Kreisberg, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

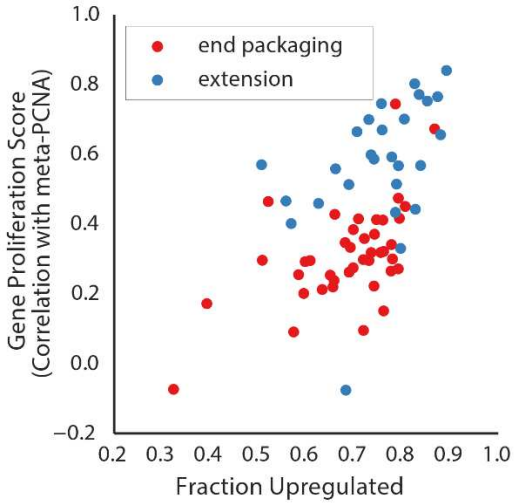
Chapter 2.7: Supplementary Figures



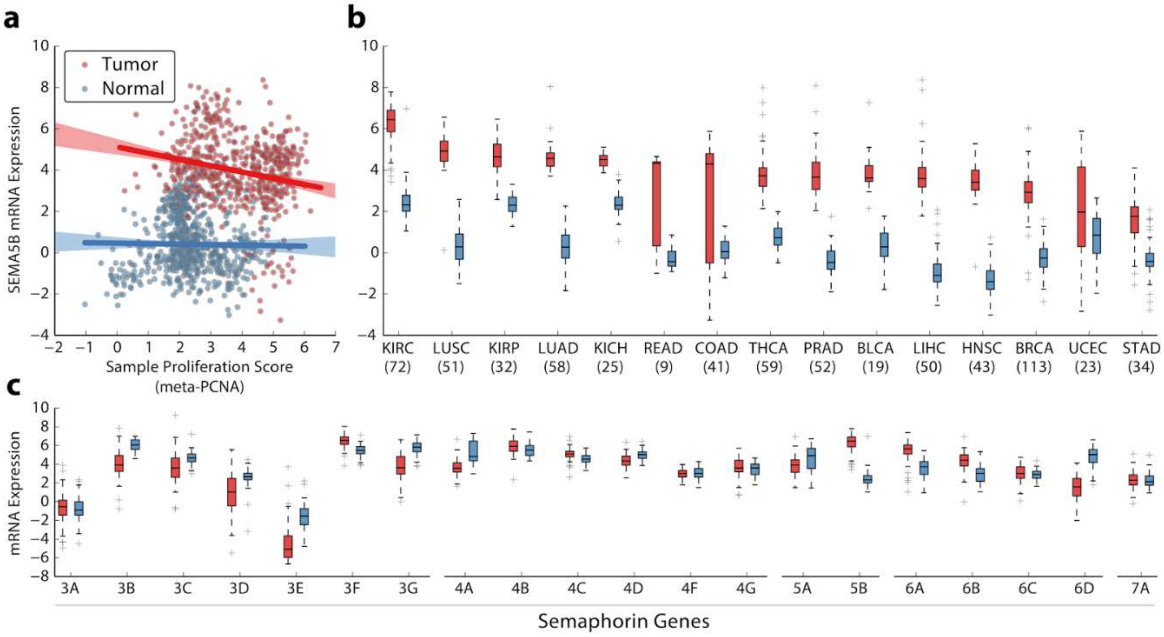
Supplementary Figure 2.1: Sample counts of TCGA patients with matched tumor/normal data.



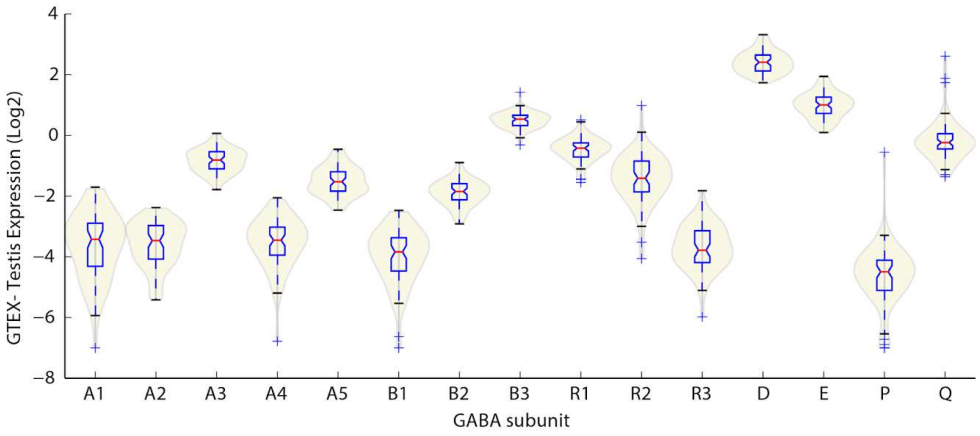
Supplementary Figure 2.2: Comparison of the fup up/down statistic to the paired t-test as an alternative metric. Shown for all genes across the pan-cancer TCGA mRNA sequencing cohort.



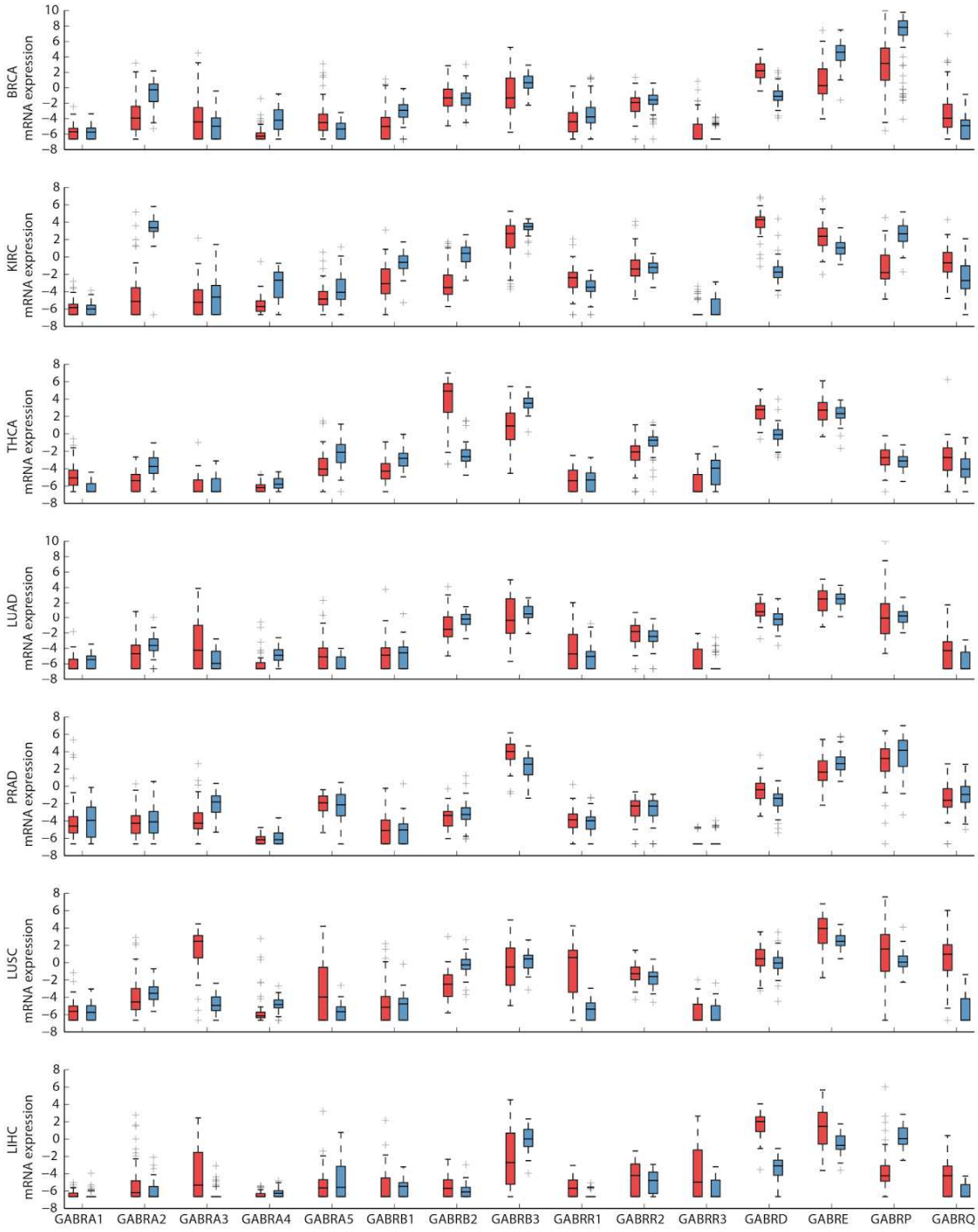
Supplementary Figure 2.3: Scatter plot comparing gene-level proliferation score against fraction upregulated for genes involved in telomere end packaging and telomere extension.



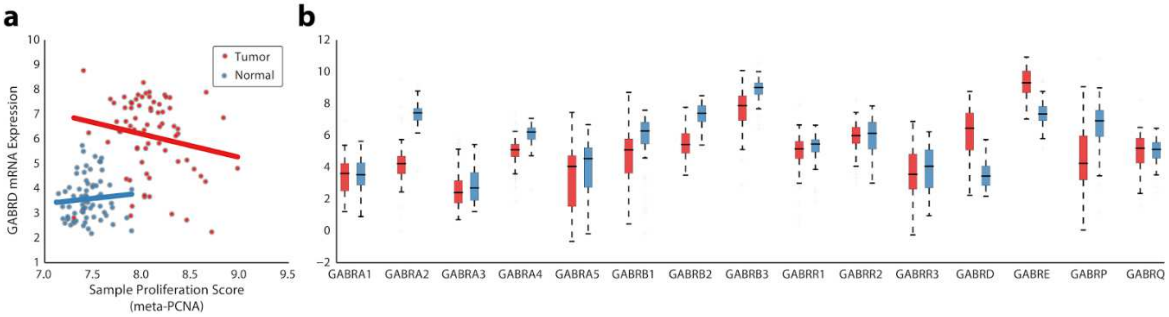
Supplementary Figure 2.4: SEMA5B is tumor-associated, independent of proliferation. (a) Scatter-plot comparing SEMA5B gene expression profiles to proliferation scores across matched tumor and normal samples. Lines indicate linear regression figs of tumor (red) and normal (blue) samples, shaded regions indicate 95% confidence intervals. (b) Comparison of matched tumor and normal profiles for SEMA5B expression, grouped by tissue type. (c) Comparison of matched tumor and normal profiles for all SEMA protein family of genes in renal cell carcinoma (note that the x-tick labels correspond to the gene suffix, e.g. 3A represents SEMA3A).



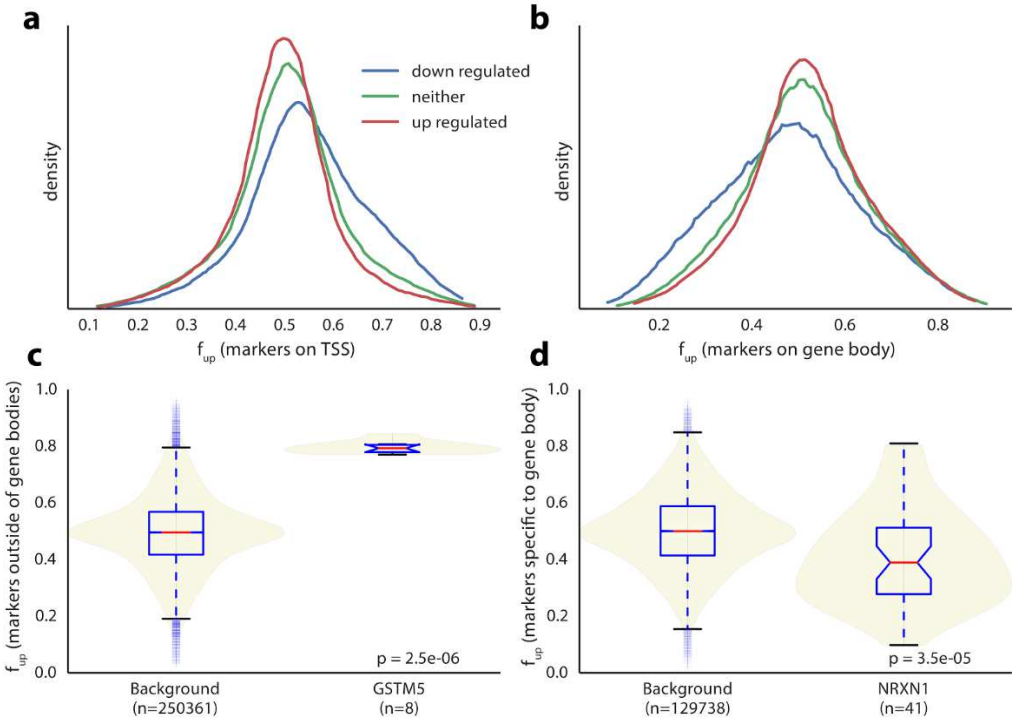
Supplementary Figure 2.5: Paired tumor-normal expression for GABA receptor genes across different tissues.



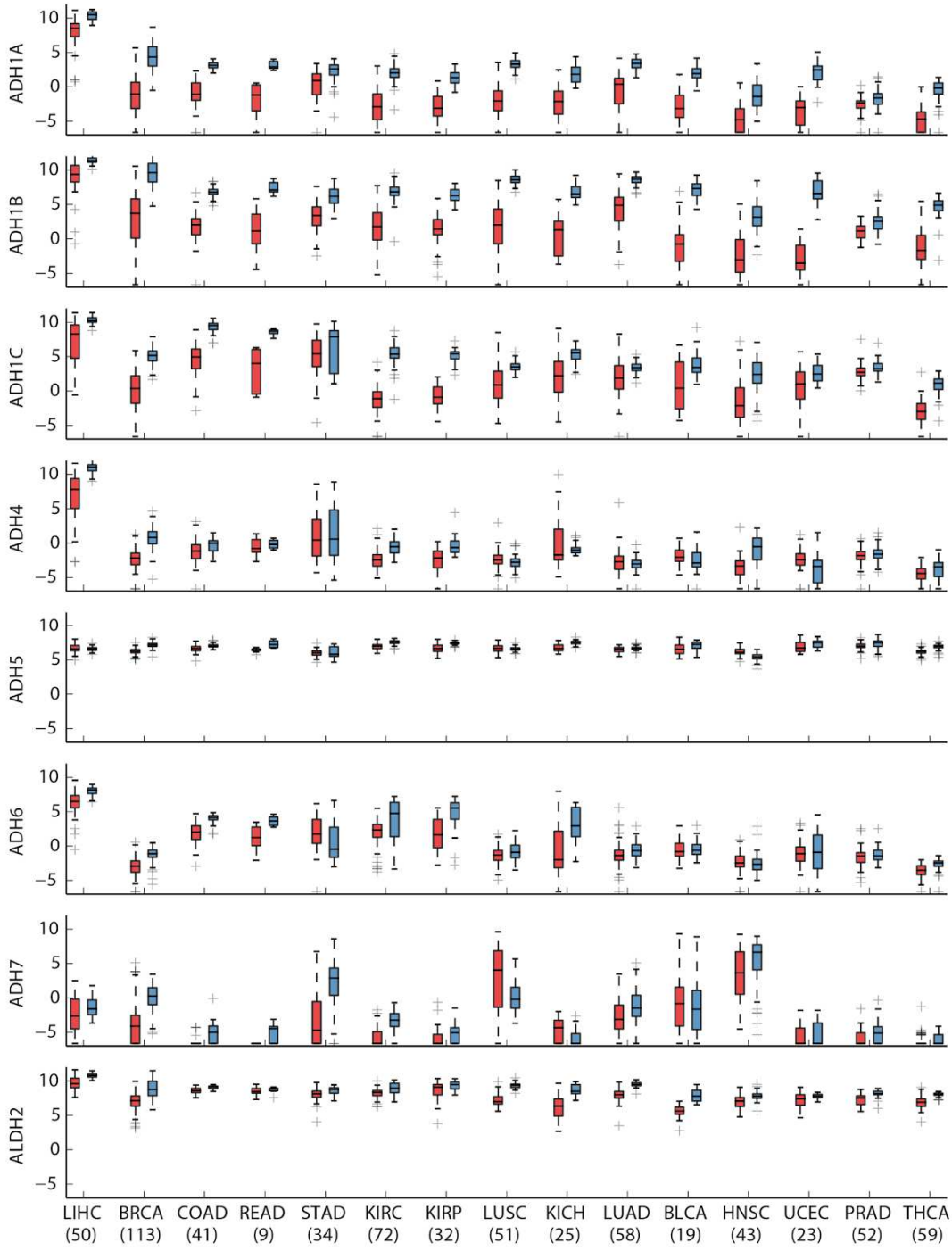
Supplementary Figure 2.6: Violin plot of of GABAA subunit gene expression in the testis. Data obtained from the Genotype-Tissue Expression (GTEX) project.



Supplementary Figure 2.7: Characterization of GABRD in a paired microarray dataset. (a) Scatter plot comparing GABRD gene expression profiles to proliferation scores across matched tumor and normal samples. (b) Comparison of matched tumor and normal profiles for all GABA protein subunits.



Supplementary Figure 2.8: Exploration of epigenetic silencing in consistently downregulated genes. (a-b) Distribution of methylation markers annotated to transcription start sites (a) or gene bodies (b), split by upregulated, downregulated or neutral status of annotated genes. Up- and down-regulation is assessed here by the significance of the detrended f_{up} metric with a threshold of $P_{Bonf} < 0.05$. (c) Comparison of probes mapping outside of the gene body on GSTM5 against similar probes annotated to all other genes. (d) Comparison of probes mapping specifically to the gene body of NRXN1 against similar probes annotated to all other genes.



Supplementary Figure 2.9: Paired tumor-normal expression for ADH genes and ALDH2.

Chapter 2.8: References

Bottos, A., Destro, E., Rissone, A., Graziano, S., Cordara, G., Assenzio, B., Cera, M.R., Mascia, L., Bussolino, F., and Arese, M. (2009). The synaptic proteins neurexins and neuroligins are widely expressed in the vascular system and contribute to its functions. Proceedings of the National Academy of Sciences 106, 20782–20787.

Broad Institute TCGA Genome Data Analysis Center (2015). Analysis-ready standardized TCGA data from Broad GDAC Firehose stddata__2015_04_02 run.

Bueno, M.J., de Castro, I.P., and Malumbres, M. (2008). Control of cell proliferation pathways by microRNAs. Cell Cycle 7, 3143–3148.

Chang, K., Creighton, C.J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y.S.N., Chu, A., Chuah, E., Chun, H.-J.E., Dhalla, N., Guin, R., Hirst, M., Hirst, C., Holt, R.A., Jones, S.J.M., Lee, D., Li, H.I., Marra, M.A., Mayo, M., Moore, R.A., Mungall, A.J., Robertson, A.G., Schein, J.E., Sipahimalani, P., Tam, A., Thiessen, N., Varhol, R.J., Beroukhim, R., Bhatt, A.S., Brooks, A.N., Cherniack, A.D., Freeman, S.S., Gabriel, S.B., Helman, E., Jung, J., Meyerson, M., Ojesina, A.I., Pedamallu, C.S., Saksena, G., Schumacher, S.E., Tabak, B., Zack, T., Lander, E.S., Bristow, C.A., Hadjipanayis, A., Haseley, P., Kucherlapati, R., Lee, S., Lee, E., Luquette, L.J., Mahadeshwar, H.S., Pantazi, A., Parfenov, M., Park, P.J., Protopopov, A., Ren, X., Santoso, N., Seidman, J., Seth, S., Song, X., Tang, J., Xi, R., Xu, A.W., Yang, L., Zeng, D., Auman, J.T., Balu, S., Buda, E., Fan, C., Hoadley, K.A., Jones, C.D., Meng, S., Mieczkowski, P.A., Parker, J.S., Perou, C.M., Roach, J., Shi, Y., Silva, G.O., Tan, D., Veluvolu, U., Waring, S., Wilkerson, M.D., Wu, J., Zhao, W., Bodenheimer, T., Hayes, D.N., Hoyle, A.P., Jeffreys, S.R., Mose, L.E., Simons, J.V., Soloway, M.G., Baylin, S.B., Berman, B.P., Bootwalla, M.S., Danilova, L., Herman, J.G., Hinoue, T., Laird, P.W., Rhie, S.K., Shen, H., Triche, T., Weisenberger, D.J., Carter, S.L., Cibulskis, K., Chin, L., Zhang, J., Getz, G., Sougnez, C., Wang, M., Saksena, G., Carter, S.L., Cibulskis, K., Chin, L., Zhang, J., Getz, G., Dinh, H., Doddapaneni, H.V., Gibbs, R., Gunaratne, P., Han, Y., Kalra, D., Kovar, C., Lewis, L., Morgan, M., Morton, D., Muzny, D., Reid, J., Xi, L., Cho, J., DiCara, D., Frazer, S., Gehlenborg, N., Heiman, D.I., Kim, J., Lawrence, M.S., Lin, P., Liu, Y., Noble, M.S., Stojanov, P., Voet, D., Zhang, H., Zou, L., Stewart, C., Bernard, B., Bressler, R., Eakin, A., Iype, L., Knijnenburg, T., Kramer, R., Kreisberg, R., Leinonen, K., Lin, J., Liu, Y., Miller, M., Reynolds, S.M., Rovira, H., Shmulevich, I., Thorsson, V., Yang, D., Zhang, W., Amin, S., Wu, C.-J., Wu, C.-C., Akbani, R., Aldape, K., Baggerly, K.A., Broom, B., Casasent, T.D., Cleland, J., Creighton, C., Dodda, D., Edgerton, M., Han, L., Herbrich, S.M., Ju, Z., Kim, H., Lerner, S., Li, J., Liang, H., Liu, W., Lorenzi, P.L., Lu, Y., Melott, J., Mills, G.B., Nguyen, L., Su, X., Verhaak,

R., Wang, W., Weinstein, J.N., Wong, A., Yang, Y., Yao, J., Yao, R., Yoshihara, K., Yuan, Y., Yung, A.K., Zhang, N., Zheng, S., Ryan, M., Kane, D.W., Aksoy, B.A., Ciriello, G., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Kahles, A., Ladanyi, M., Lee, W., Lehmann, K.-V., Miller, M.L., Ramirez, R., Rätsch, G., Reva, B., Sander, C., Schultz, N., Senbabaoglu, Y., Shen, R., Sinha, R., Sumer, S.O., Sun, Y., Taylor, B.S., Weinhold, N., Fei, S., Spellman, P., Benz, C., Carlin, D., Cline, M., Craft, B., Ellrott, K., Goldman, M., Haussler, D., Ma, S., Ng, S., Paull, E., Radenbaugh, A., Salama, S., Sokolov, A., Stuart, J.M., Swatloski, T., Uzunangelov, V., Waltman, P., Yau, C., Zhu, J., Hamilton, S.R., Getz, G., Sougnez, C., Abbott, S., Abbott, R., Dees, N.D., Delehaunty, K., Ding, L., Dooling, D.J., Eldred, J.M., Fronick, C.C., Fulton, R., Fulton, L.L., Kalicki-Veizer, J., Kanchi, K.-L., Kandoth, C., Koboldt, D.C., Larson, D.E., Ley, T.J., Lin, L., Lu, C., Magrini, V.J., Mardis, E.R., McLellan, M.D., McMichael, J.F., Miller, C.A., O'Laughlin, M., Pohl, C., Schmidt, H., Smith, S.M., Walker, J., Wallis, J.W., Wendl, M.C., Wilson, R.K., Wylie, T., Zhang, Q., Burton, R., Jensen, M.A., Kahn, A., Pihl, T., Pot, D., Wan, Y., Levine, D.A., Black, A.D., Bowen, J., Frick, J., Gastier-Foster, J.M., Harper, H.A., Helsel, C., Leraas, K.M., Lichtenberg, T.M., McAllister, C., Ramirez, N.C., Sharpe, S., Wise, L., Zmuda, E., Chanock, S.J., Davidsen, T., Demchok, J.A., Eley, G., Felau, I., Ozenberger, B.A., Sheth, M., Sofia, H., Staudt, L., Tarnuzzer, R., Wang, Z., Yang, L., Zhang, J., Omberg, L., Margolin, A., Raphael, B.J., Vandin, F., Wu, H.-T., Leiserson, M.D.M., Benz, S.C., Vaske, C.J., Noushmehr, H., Knijnenburg, T., Wolf, D., Veer, L.V., Collisson, E.A., Anastassiou, D., Yang, T.-H.O., Lopez-Bigas, N., Gonzalez-Perez, A., Tamborero, D., Xia, Z., Li, W., Cho, D.-Y., Przytycka, T., Hamilton, M., McGuire, S., Nelander, S., Johansson, P., Jörnsten, R., Kling, T., Sanchez, J., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics 45, 1113–1120.

Cheng, W.-Y., Yang, T.-H.O., and Anastassiou, D. (2013). Biomolecular Events in Cancer Revealed by Attractor Metagenes. PLoS Computational Biology 9, e1002920.

Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., Fulton, L., Fulton, R.S., Zhang, Q., Wendl, M.C., Lawrence, M.S., Larson, D.E., Chen, K., Dooling, D.J., Sabo, A., Hawes, A.C., Shen, H., Jhangiani, S.N., Lewis, L.R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S.E., Clerc, K., Metcalf, G.A., Ng, B., Milosavljevic, A., Gonzalez-Garay, M.L., Osborne, J.R., Meyer, R., Shi, X., Tang, Y., Koboldt, D.C., Lin, L., Abbott, R., Miner, T.L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B.H., Kraja, A., Crosby, S.D., Sawyer, C.S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L.R., Dutt, A., Fennell, T., Hanna, M., Johnson, B.E., Onofrio, R.C., Thomas, R.K., Tonon, G.,

Weir, B.A., Zhao, X., Ziaugra, L., Zody, M.C., Giordano, T., Orringer, M.B., Roth, J.A., Spitz, M.R., Wistuba, I.I., Ozenberger, B., Good, P.J., Chang, A.C., Beer, D.G., Watson, M.A., Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W.D., Pao, W., Province, M.A., Weinstock, G.M., Varmus, H.E., Gabriel, S.B., Lander, E.S., Gibbs, R.A., Meyerson, M., and Wilson, R.K. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455, 1069–1075.

Evan, G.I., and Vousden, K.H. (2001). Proliferation, cell cycle and apoptosis in cancer. Nature 411, 342–348.

Gardina, P.J., Clark, T.A., Shimada, B., Staples, M.K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., Davies, C., Williams, A., and Turpaz, Y. (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics 7, 325.

Geigerseder, C., Doepner, R., Thalhammer, A., Frungieri, M.B., Gamel-Didelon, K., Calandra, R.S., Köhn, F.M., and Mayerhofer, A. (2003). Evidence for a GABAergic system in rodent and human testis: local GABA production and GABA receptors. Neuroendocrinology 77, 314–323.

Gentles, A.J., Newman, A.M., Liu, C.L., Bratman, S.V., Feng, W., Kim, D., Nair, V.S., Xu, Y., Khuong, A., Hoang, C.D., Diehn, M., West, R.B., Plevritis, S.K., and Alizadeh, A.A. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. Nature Medicine 21, 938–945.

Hamfjord, J., Stangeland, A.M., Hughes, T., Skrede, M.L., Tveit, K.M., Ikdahl, T., and Kure, E.H. (2012). Differential Expression of miRNAs in Colorectal Cancer: Comparison of Paired Tumor Tissue and Adjacent Normal Mucosa Using High-Throughput Sequencing. PLoS ONE 7, e34150.

Kobayashi, Y., Absher, D.M., Gulzar, Z.G., Young, S.R., McKenney, J.K., Peehl, D.M., Brooks, J.D., Myers, R.M., and Sherlock, G. (2011). DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer. Genome Research 21, 1017–1027.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., Kiezun, A., Hammerman, P.S., McKenna, A., Drier, Y., Zou, L., Ramos, A.H., Pugh, T.J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M.L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D.I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A.M., Lohr, J., Landau, D.-A., Wu, C.J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S.A., Mora, J., Lee, R.S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S.B., Roberts, C.W.M., Biegel, J.A., Stegmaier, K., Bass, A.J., Garraway, L.A.,

Meyerson, M., Golub, T.R., Gordenin, D.A., Sunyaev, S., Lander, E.S., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–218.

Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature 469, 343–349.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E.G., Brat, D.J., M. Mastrogianakis, G., Olson, J.J., Mikkelsen, T., Lehman, N., Aldape, K., Alfred Yung, W.K., Bogler, O., VandenBerg, S., Berger, M., Prados, M., Muzny, D., Morgan, M., Scherer, S., Sabo, A., Nazareth, L., Lewis, L., Hall, O., Zhu, Y., Ren, Y., Alvi, O., Yao, J., Hawes, A., Jhangiani, S., Fowler, G., San Lucas, A., Kovar, C., Cree, A., Dinh, H., Santibanez, J., Joshi, V., Gonzalez-Garay, M.L., Miller, C.A., Milosavljevic, A., Donehower, L., Wheeler, D.A., Gibbs, R.A., Cibulskis, K., Sougnez, C., Fennell, T., Mahan, S., Wilkinson, J., Ziaugra, L., Onofrio, R., Bloom, T., Nicol, R., Ardlie, K., Baldwin, J., Gabriel, S., Lander, E.S., Ding, L., Fulton, R.S., McLellan, M.D., Wallis, J., Larson, D.E., Shi, X., Abbott, R., Fulton, L., Chen, K., Koboldt, D.C., Wendl, M.C., Meyer, R., Tang, Y., Lin, L., Osborne, J.R., Dunford-Shore, B.H., Miner, T.L., Delehaunty, K., Markovic, C., Swift, G., Courtney, W., Pohl, C., Abbott, S., Hawkins, A., Leong, S., Haipek, C., Schmidt, H., Wiechert, M., Vickery, T., Scott, S., Dooling, D.J., Chinwalla, A., Weinstock, G.M., Mardis, E.R., Wilson, R.K., Getz, G., Winckler, W., Verhaak, R.G.W., Lawrence, M.S., O'Kelly, M., Robinson, J., Alexe, G., Beroukhim, R., Carter, S., Chiang, D., Gould, J., Gupta, S., Korn, J., Mermel, C., Mesirov, J., Monti, S., Nguyen, H., Parkin, M., Reich, M., Stransky, N., Weir, B.A., Garraway, L., Golub, T., Meyerson, M., Chin, L., Protopopov, A., Zhang, J., Perna, I., Aronson, S., Sathiamoorthy, N., Ren, G., Yao, J., Wiedemeyer, W.R., Kim, H., Won Kong, S., Xiao, Y., Kohane, I.S., Seidman, J., Park, P.J., Kucherlapati, R., Laird, P.W., Cope, L., Herman, J.G., Weisenberger, D.J., Pan, F., Van Den Berg, D., Van Neste, L., Mi Yi, J., Schuebel, K.E., Baylin, S.B., Absher, D.M., Li, J.Z., Southwick, A., Brady, S., Aggarwal, A., Chung, T., Sherlock, G., Brooks, J.D., Myers, R.M., Spellman, P.T., Purdom, E., Jakkula, L.R., Lapuk, A.V., Marr, H., Dorton, S., Gi Choi, Y., Han, J., Ray, A., Wang, V., Durinck, S., Robinson, M., Wang, N.J., Vranizan, K., Peng, V., Van Name, E., Fontenay, G.V., Ngai, J., Conboy, J.G., Parvin, B., Feiler, H.S., Speed, T.P., Gray, J.W., Brennan, C., Socci, N.D., Olshen, A., Taylor, B.S., Lash, A., Schultz, N., Reva, B., Antipin, Y., Stukalov, A., Gross, B., Cerami, E., Qing Wang, W., Qin, L.-X., Seshan, V.E., Villafania, L., Cavatore, M., Borsu, L., Viale, A., Gerald, W., Sander, C., Ladanyi, M., Perou, C.M., Neil Hayes, D., Topal, M.D., Hoadley, K.A., Qi, Y., Balu, S., Shi, Y., Wu, J., Penny, R., Bittner, M., Shelton, T., Lenkiewicz, E., Morris, S., Beasley, D., Sanders, S., Kahn, A., Sfeir, R., Chen, J., Nassau, D., Feng, L., Hickey, E., Zhang, J., Weinstein, J.N., Barker, A., Gerhard, D.S., Vockley, J., Compton, C., Vaught, J., Fielding, P., Ferguson, M.L., Schaefer, C., Madhavan, S., Buetow, K.H., Collins, F., Good, P., Guyer, M., Ozenberger, B.,

Peterson, J., and Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–1068.

Mele, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., Johnson, R., Segre, A.V., Djebali, S., Niarchou, A., Consortium, T.G., Wright, F.A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E.T., Ardlie, K.G., and Guigo, R. (2015). The human transcriptome across tissues and individuals. Science 348, 660–665.

Notterman, D.A., Alon, U., Sierk, A.J., and Levine, A.J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res. 61, 3124–3130.

Nusser, Z., Roberts, J.D., Baude, A., Richards, J.G., and Somogyi, P. (1995). Relative densities of synaptic and extrasynaptic GABAA receptors on cerebellar granule cells as determined by a quantitative immunogold method. J. Neurosci. 15, 2948–2960.

Perez, F., and Granger, B.E. (2007). IPython: A System for Interactive Scientific Computing. Computing in Science & Engineering 9, 21–29.

Rahman, M., Jackson, L.K., Johnson, W.E., Y. Li, D., Bild, A.H., and Piccolo, S.R. (2015). Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. Bioinformatics btv377.

Schell, J.C., Olson, K.A., Jiang, L., Hawkins, A.J., Van Vranken, J.G., Xie, J., Egnatchik, R.A., Earl, E.G., DeBerardinis, R.J., and Rutter, J. (2014). A Role for the Mitochondrial Pyruvate Carrier as a Repressor of the Warburg Effect and Colon Cancer Cell Growth. Molecular Cell 56, 400–413.

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. Nature Genetics 36, 1090–1098.

Seo, J.-S., Ju, Y.S., Lee, W.-C., Shin, J.-Y., Lee, J.K., Bleazard, T., Lee, J., Jung, Y.J., Kim, J.-O., Shin, J.-Y., Yu, S.-B., Kim, J., Lee, E.-R., Kang, C.-H., Park, I.-K., Rhee, H., Lee, S.-H., Kim, J.-I., Kang, J.-H., and Kim, Y.T. (2012). The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Research 22, 2109–2119.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545–15550.

Tamagnone, L. (2012). Emerging Role of Semaphorins as Major Regulatory Signals and Potential Therapeutic Targets in Cancer. Cancer Cell 22, 145–152.

Terunuma, A., Putluri, N., Mishra, P., Mathé, E.A., Dorsey, T.H., Yi, M., Wallace, T.A., Issaq, H.J., Zhou, M., Killian, J.K., Stevenson, H.S., Karoly, E.D., Chan, K., Samanta, S., Prieto, D., Hsu, T.Y.T., Kurley, S.J., Putluri, V., Sonavane, R., Edelman, D.C., Wulff, J., Starks, A.M., Yang, Y., Kittles, R.A., Yfantis, H.G., Lee, D.H., Ioffe, O.B., Schiff, R., Stephens, R.M., Meltzer, P.S., Veenstra, T.D., Westbrook, T.F., Sreekumar, A., and Ambs, S. (2014). MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. Journal of Clinical Investigation 124, 398–412.

Tian, J., Lu, Y., Zhang, H., Chau, C.H., Dang, H.N., and Kaufman, D.L. (2004). Gamma-aminobutyric acid inhibits T cell autoimmunity and the development of inflammatory responses in a mouse type 1 diabetes model. J. Immunol. 173, 5298–5304.

Venet, D., Dumont, J.E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput. Biol. 7, e1002240.

Warburg, O. (1956). On the origin of cancer cells. Science 123, 309–314.

Wierstra, I., and Alves, J. (2007). FOXM1, a typical proliferation-associated transcription factor. Biol. Chem. 388, 1257–1274.

Young, S.Z., and Bordey, A. (2009). GABA's Control of Stem and Cancer Cell Proliferation in Adult Neural and Peripheral Niches. Physiology 24, 171–185.

Chapter 3: Methylome-wide analysis of chronic HIV infection reveals five-year increase in biological age and epigenetic targeting of HLA

Chapter 3.1: Highlights

- Methylome-wide analysis of HIV chronically infected, cART treated individuals
- HIV increases epigenetic aging by 4.9 years, associated with expected 19% increased mortality
- HLA locus is hypomethylated in HIV+ individuals
- Development and multi-cohort validation of epigenetic clock models of aging that are independent of sample cell type composition

Chapter 3.2: Summary

HIV-infected individuals are living longer on antiretroviral therapy, but many patients display signs that in some ways resemble premature aging. To investigate and quantify the impact of chronic HIV infection on aging, we report a global analysis of the whole blood DNA methylomes of 137 HIV+ individuals under sustained therapy along with 44 matched HIV− individuals and 1,200 population controls. First, we develop and validate epigenetic models of aging that are independent of blood cell composition. Using this technique, we find that both chronic and recent HIV infection lead to an average aging advancement of 4.9 years, increasing expected mortality risk by 19%. In addition, sustained infection results in global deregulation of the methylome across >80,000 CpGs and specific

hypomethylation of the genomic region encoding the human leukocyte antigen locus (HLA). We find that decreased HLA methylation is predictive of lower CD4 / CD8 T cell ratio, linking molecular aging, epigenetic regulation and disease progression.

Chapter 3.3: Introduction

It is an open question why some people show early or delayed onset of age-associated disorders (Kennedy *et al.*, 2014). Recent studies have found that aging is associated with epigenetic changes (Christensen *et al.*, 2009; Day *et al.*, 2013; Heyn *et al.*, 2012; Maegawa *et al.*, 2010; Numata *et al.*, 2012; Rakyan *et al.*, 2010; West *et al.*, 2013), and based on this work we (Hannum *et al.*, 2012) and others (Horvath, 2013; Weidner *et al.*, 2014) have built models capable of predicting a person's age using DNA methylation patterns across a large number of CpG sites. Although these models are fairly accurate, errors of prediction — differences between the chronological and predicted age — serve as a quantitative readout of the relative advancement or retardation of the 'biological age' of an individual. Biological age advancement has been correlated with factors such as gender, genetic polymorphisms and diseases including cancer and diabetes, and it may influence the onset of other age-associated disorders (Day *et al.*, 2013; Hannum *et al.*, 2012). A recent longitudinal study has validated the clinical utility of these models by demonstrating a link between biological age advancement and increased mortality rates (Marioni *et al.*, 2015).

Biological aging has become of particular interest in treatment of HIV, in which the development of combination active anti-retroviral therapy (cART) now enables infected individuals to live many decades (Deeks, 2011; Deeks *et al.*, 2013; Maartens *et al.*, 2014). Several studies have suggested links between chronic HIV infection and early onset of neurodegeneration (Nightingale *et al.*, 2014), liver or kidney failure (Hilton, 2013; Joshi *et al.*, 2011; Kovari *et al.*, 2013), cancer (Dubrow *et al.*, 2012), cardiovascular disease (Freiberg *et al.*, 2013), or telomere shortening (Chou *et al.*, 2013; Leeansyah *et al.*, 2013; Pathai *et al.*, 2013), leading to the hypothesis that HIV+ patients might experience advanced or accelerated aging (Appay and Rowland-Jones, 2002; Guaraldi *et al.*, 2011; Smith *et al.*, 2012). While these studies report rough estimates of HIV-mediated age advancement in the range of 0-20 years, it has been difficult to accurately quantify this number due to sampling effects, co-morbidities, and relatively low incidence rates of any single age-associated disease. To this effect, the existence, extent, and molecular basis of a bona-fide increase in aging have been unclear (Althoff *et al.*, 2014; Solomon *et al.*, 2014), in part due to lack of an objective biological clock or aging biomarker.

In parallel with such epidemiological observations, a number of studies report age effects using blood-based biomarkers. Analysis of cell surface markers in T cells has shown HIV+ subjects to show phenotypes of older cells (Cao *et al.*, 2009). Other studies have observed shortened telomeres in certain cell populations (Rickabaugh *et al.*, 2011) as well as whole blood (Zanet *et al.*, 2014), indirectly linking HIV to aging via the well-studied connection between telomere

length and age (Lindsey *et al.*, 1991; Cawthon *et al*, 2003). Furthermore, a recent analysis of untreated HIV+ individuals found DNA methylation sites that are associated with both HIV infection and age (Rickabaugh *et al.*, 2014). Together, these results raise the possibility that HIV infection results in an increase in biological age. Many questions remain, however: Are the epigenetic changes associated with HIV the same as those previously identified (Hannum *et al.*, 2012; Horvath, 2013) in normal individuals as markers of 'biological age', and how complete is the correspondence between these two responses? What is the quantitative effect on aging in years, and is it fixed age advancement or continuous acceleration? What is the impact on aging of chronic HIV infection and sustained cART treatment? Are there other impacts of HIV on the methylome that are unrelated to aging?

Here we begin to address these questions by analyzing the methylomes of HIV-infected, cART-treated subjects, in which we observe a strong shared phenotype of HIV and age. To understand this signal, we develop models of biological age that allow us to establish a clear quantitative link between HIV infection and aging as observed in the general population. We identify both global and targeted epigenomic effects of HIV, including specific hypomethylation of the HLA locus. Together, these results shed light on the epigenetic consequences and gerontological aspects of chronic HIV infection.

Chapter 3.4: Results

Chapter 3.4.1: Genome-wide DNA methylation profiling

To determine whether HIV is associated with signs of aberrant biological aging, samples of whole blood DNA were obtained from 137 HIV-infected, cART-treated but otherwise healthy non-Hispanic white males (no hepatitis C co-infection, no diabetes, and high adherence to therapy) and 44 healthy non-Hispanic white male controls (**Supplementary Table 3.1, Supplementary Figure 3.1**). Genome-wide methylation profiles of each sample were determined using the Illumina Infinium HumanMethylation450 BeadChip array. Data were normalized and controlled for quality using standard techniques, resulting in removal of two control patients due to poor signal (**Experimental Procedures**).

Chapter 3.4.2: Unsupervised analysis shows shared phenotypes of HIV and age

In preliminary data exploration, we performed an unsupervised analysis to identify age-associated methylation sites and explore their relation to HIV infection. Analysis of a previous methylome-wide screen of 538 healthy subjects (Hannum *et al.*, 2012) identified 61,592 methylation sites associated with age at a 1% false-discovery rate (FDR, likelihood ratio test in multivariate regression model with Benjamini-Hochberg correction). Validation of these sites in whole blood from a second control cohort from the European Prospective Investigation into Cancer and Nutrition (Riboli *et al.*, 2002) (EPIC, N = 662) confirmed 26,927 sites as strongly associated with age (**Figure 3.1A, Supplementary Table 3.2**).

Among these validated age-associated sites, we found a striking association with methylation in the HIV+ patients relative to healthy controls (P <

10^{-100}, **Figure 3.1B**). Further analysis of these sites found a positive association of the first principal component with both age and HIV status (**Figure 3.1C**, **Supplementary Table 3.3**, association by multivariate linear model $P < 10^{-8}$). These findings support a link between HIV infection and aging (Rickabaugh *et al.*, 2014), as quantitatively measured by epigenomic profiling (**Figure 3.1D**).

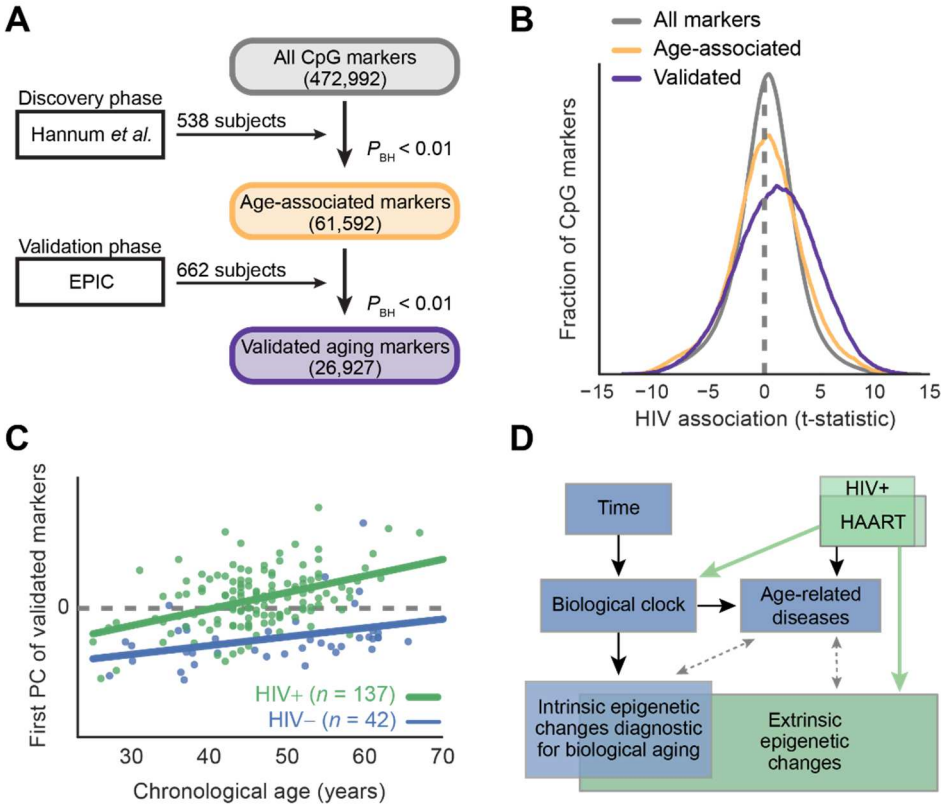


Figure 3.1: Shared epigenetic signature of HIV infection and aging. A, Discovery and validation of CpG methylation markers associated with age. **B**, Distribution of t-statistics measuring association of each methylation marker with HIV status. Colors indicate groups of markers identified in (**A**): Gray, all markers; yellow, age-associated markers from discovery phase; violet, subset of age-associated markers confirmed in validation. **C**, Principal component (PC) analysis of the validated age-associated markers, in which the first PC (y-axis) is positively associated with both age (x-axis) and disease status (HIV+, green; HIV–, blue). **D**, Potential relationships among HIV infection, epigenetics, disease, and aging. Black: known; Dashed gray: potential; Green: connections explored in this study.

Chapter 3.4.3: Benchmarking and refinement of epigenetic aging models

Given the shared effects of HIV and aging, we sought to determine whether HIV causes the same biological aging signature as previously found in cohorts of uninfected individuals (Hannum *et al.*, 2012; Horvath *et al.*, 2013, Marioni *et al.*, 2015). We tested aging models from both our group (Hannum *et al.*, 2012) and Horvath (Horvath, 2013) in three independent datasets derived from whole blood samples (Hannum *et al.*, 2012; Riboli *et al.*, 2002; this study; **Supplementary Table 3.2**). Although the Hannum and Horvath modeling efforts were based on different methodologies and training data, we found they made very similar predictions (r = 0.9, Pearson's correlation, **Figure 3.2A**), and furthermore found that a consensus of the two models outperformed either model individually (**Figures 3.2B and 3.2C, Supplementary Table 3.2**). For this reason, we used this consensus model for all analyses throughout our study.

A potential issue with these models arises in the fact that methylation profiles from whole blood are influenced by cell composition, and different cell types have different methylation states (Jaffe and Irizarry, 2014). These differences might be particularly pronounced in HIV-infected patients, some of whom have low $CD4^+$ T cell counts (Trono *et al.*, 2010). To understand the sensitivity of epigenetic aging models to cell type composition, we obtained datasets with methylation profiles derived from sorted populations of blood (Houseman *et al.*, 2012; Absher *et al.*, 2013; Reynolds *et al.*, 2014). Examination of the Houseman *et al.* dataset, consisting of methylome profiles of nine sorted blood cell types across six individuals, indicated that epigenetic age

measurements for each cell type were concordant with whole-blood measurements in those same patients (**Supplementary Figure 3.2**). To further understand the reproducibility of epigenetic age in such purified cell populations, we downloaded two datasets profiling sorted cells across shared sets of individuals (Absher *et al.*, 2013, GSE59250; Reynolds *et al.*, 2014, GSE56046). Among these sorted cell datasets, we saw good concordance of epigenetic age predictions with chronological age (**Supplementary Figure 3. 3A-F**). Epigenetic age was reproducible across different cell types profiled from the same patients, with high agreement of age estimates (r > 0.77-0.88) and moderate but very significant agreement of age advancement (Pearson's r > 0.45-0.68; P < 0.0001 for all associations, **Supplementary Figure 3.3G-J**).

While we expect the contribution of cell composition to be minimal, we nonetheless developed an algorithm to individually normalize each methylation profile using the methylation-derived cell type information. In brief, we use using a previously reported method (Jaffe and Irizarry, 2014) to reliably predict blood composition (**Supplementary Figure 3.4**), and adjust out the expected contribution of cell-type specific effects. This procedure greatly limited the effects of age and HIV induced blood composition changes in downstream analyses (**Experimental Procedures**, **Supplementary Figure 3.5**).

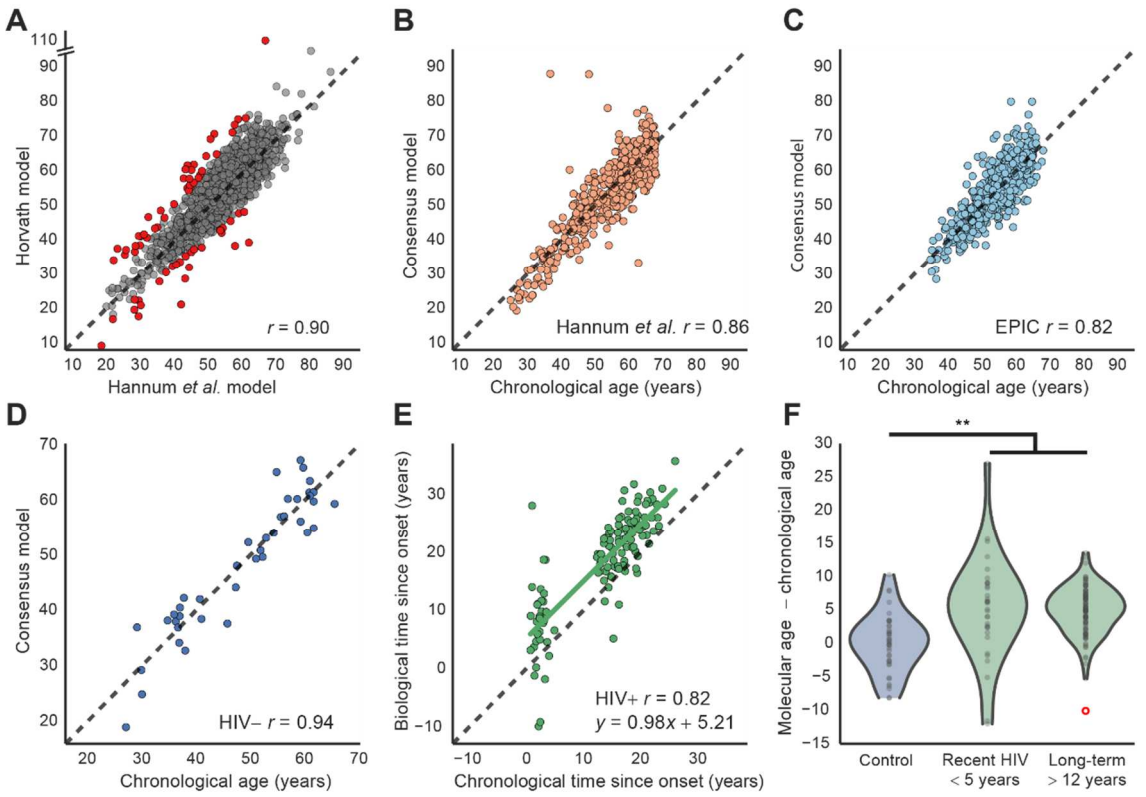


Figure 3.2: Epigenetic models accurately predict age and indicate advanced aging for HIV-infected individuals. A, Scatter plot comparing the ages predicted using the Hannum *et al.* and Horvath models on healthy controls (n = 1,246 from HIV–, Hannum *et al.* and EPIC datasets). Red points indicate patients that were discarded due to disagreement between the two aging models (n = 66). **B-C**, Accuracy of the consensus model (y-axis) to predict true chronological age (x-axis) in datasets from Hannum *et al*. (n = 497, **B**) or EPIC (n = 637, **C**). Panels (**A-C**) show patients between 25 and 68 years old. **D**, Scatter plot of predicted biological age (consensus aging model) *versus* chronological age for HIV– healthy controls. **E**, Scatter plot of biological time *versus* chronological time since HIV onset for infected subjects. **F,** Violin plots showing the distribution of residuals from regression of biological *versus* chronological age. Three groups are shown: HIV– controls, short-term HIV+ infected individuals, and long-term HIV+ infected individuals. Note that the red circle indicates an outlier, which is not used to fit the violin profile, but is used in all statistical assessments. **A-E**, Black dashed lines indicate diagonal ($y = x$). r, Pearson's correlation coefficient. ** indicates $P < 10^{-5}$.

Chapter 3.4.4: HIV+ individuals have advanced DNA methylation age

We next used this consensus aging model to calculate the 'biological age' of each individual in our cohort (**Supplementary Table 3.2**). For uninfected controls, the calculated biological age had a very high concordance with

chronological age (**Figure 3.2D**, Pearson's $r = 0.94$). In contrast, the HIV+ patients had a biological age advancement of 4.9 years on average ($P < 10^{-8}$ by Student's t-test, 95% confidence interval 3.4 - 7.1 years, **Figures 3.2E and 3.2F**). These results were consistent with our previous unsupervised analysis (**Figures 3.1B and 3.1C**) in suggesting that HIV infection leads to advanced instead of accelerated aging. Furthermore, we found that the age advancement of HIV+ individuals was negatively correlated with the ratio of $CD4^+$ / $CD8^+$ T lymphocytes (Spearman's rho = -0.2, $P < 0.02$). $CD4^+$ T cells are a major indicator of immune integrity (Leung *et al.*, 2013; Serrano-Villar *et al.*, 2014) and are inversely associated with morbidity and mortality, including from non-AIDS defining diseases (SMART Study Group *et al.*, 2006); similarly, the CD4/CD8 ratio predicts non-AIDS morbidity (Leung *et al.*, 2013; Serrano-Villar *et al.*, 2014). This finding links biological aging of HIV infected individuals to a clinical measure of disease progression, and it raises the possibility that patients with stable immune responses may be less affected by the advanced aging phenotype. Taking into account a recent of 4.2% increase in mortality risk per year of biological age advancement using the Hannum model estimate (Marioni *et al.*, 2015), the changes observed in HIV+ patients result in an expected total mortality risk increase of 19%.

Chapter 3.4.5: Age advancement is independent of HIV duration

Notably, patients more recently infected with HIV (< 5 years) had no significant difference in age advancement from those patients with chronic (>12 years) infection ($P > 0.5$, Mann-Whitney U Test; **Figure 3.2F**). Similar findings

emerged from a regression analysis of the chronological *versus* biological time since infection: the slope did not differ from one (0.98 ± 0.06, standard error) whereas the y-intersect was significantly positive (5.2 ± 0.9; **Figures 3.2E and 3.2F**). These findings lend support to the theory that age advancement occurs early in the course of disease as a consequence of acute infection or reaction to drug treatment (Guaraldi *et al.*, 2011; Smith *et al.*, 2012). The lack of an increase of age advancement with disease duration seems to contradict alternative views that HIV-mediated aging occurs through cumulative effects of latent virus (Appay and Rowland-Jones, 2002) or chronic therapeutic intervention (Torres and Lewis, 2014). We did however observe less variation in age advancement within the chronically infected HIV+ individuals (**Figure 3.2F**, $P < 0.002$, Bartlett's test relative to recently infected group), perhaps reflecting the comparative stability of infection and immune response on long-term cART therapy (Luz *et al.*, 2014; Rosenblatt *et al.*, 2005).

Chapter 3.4.6: Age advancement is independent of cellular composition

While the direct effects of cell type composition on the whole blood methylome were corrected by the adjustment described in the experimental procedures, we considered that it was still possible that changes in cell type composition could lead to downstream, indirect changes in the epigenomes of all blood cells. If this were the case, cell-type associated changes could be responsible for the observed increase in biological age in the HIV+ cases. To assess this possibility, we constructed a multivariate linear model in which cell type composition variables and HIV status were used to predict biological age as

measured by the methylome (**Table 3.1**). In this model, the presence of HIV was associated with an age advancement of 3.8±1.1 years, while the presence of natural killer cells accounted for additional increases in biological age (**Table 3.1A**). In an even more conservative test, we modeled age advancement with cell type composition variables alone and found that the unexplained variation in this model still had a significant association with HIV infection (P = 0.02, Likelihood Ratio Test, **Table 3.1B**). Thus, even in a very conservative analysis, HIV infection still has association with advanced aging that is entirely independent of cell composition.

We then sought to experimentally assess if the observed age advancement due to HIV infection was also observed in purified cell populations. Using standard calculations of statistical power, we estimated that a sample of 48 patients, balanced approximately between cases and controls, would have 81% power to detect the same aging advancement effect as our primary screen at $p < 0.01$. Accordingly, this number of subjects was prospectively recruited from the University of Pittsburg Medical Center (**Experimental Procedures**, **Supplementary Table 3.4**). Whole blood was separated with flow cytometry to isolate pure populations of neutrophils and $CD4^+$ T-cells.

Table 3.1: Multivariate linear models of biological age based on chronological age, HIV and cellular composition.

A. One-Step Model

A1. Dependent variable: Biological age

	Independent variable	Effect	StdErr	*t*	*P*
	HIV	3.76	1.14	3.3	0.003
	Chronological age	–0.12	0.04	–3.0	0.001
Cell composition (%)	NK cell	0.21	0.08	2.6	0.011
	CD4 T cell	–0.07	0.07	–1.1	0.293
	CD8 T cell	0.13	0.06	0.2	0.812
	B cell	–0.17	0.13	–1.4	0.174
	Monocyte	–0.09	0.14	–0.6	0.521

B. Two-Step Model†

B1. Dependent variable: Biological age

	Independent variable	Effect	StdErr	*t*	*P*
	Chronological age	–0.22	0.07	–3.1	0.002
Cell composition (%)	NK cell	0.15	0.07	2.1	0.041
	CD4 T cell	–0.23	0.08	–2.9	0.004
	CD8 T cell	0.15	0.08	1.8	0.076
	B cell	–0.10	0.07	–1.3	0.190
	Monocyte	–0.05	0.07	–0.7	0.492

B2. Dependent variable: Step B1 model residuals

	HIV	0.18	0.08	2.4	0.019

† In the two step model **(B)**, residuals from the model using chronological age and cellular composition are carried over to a second regression using HIV status.

As in whole blood, unsupervised analysis showed a clear effect of HIV in age associated methylation markers (**Figure 3.3A-B**). Application of epigenetic models of aging in these pure-cell data-sets showed good concordance of predicted age with chronological age in both cell types (**Figure 3.3C-F**).

Interestingly, while the two aging models had fairly good agreement in age advancement across the cohort, the Hannum model predicted a 2.5 year increase in age ($P < 0.03$, 95% CI 0.6-5.0 years, **Figure 3.3E**) whereas the Horvath model had a much smaller effect of 0.4 year ($P > 0.05$). In contrast, $CD4^+$ T-cells had a much stronger and more consistent signal across the models with a consensus aging model showing an increase of 5.7 years in the HIV+ subjects ($P < 10^{-5}$, 95% CI 3.4-7.9 years, **Figure 3.3F**). These data show that the effect of epigenetic age advancement is not merely an artifact of changing blood composition, but rather is likely to reflect true aging signals. The stronger effect size within $CD4^+$ T-cells (**Figure 3.3G-H**) suggests that these cells may be suffer from more age-like stress than neutrophils, and further work is likely needed to understand how disease may affect aging rates across different cell-types and tissues.

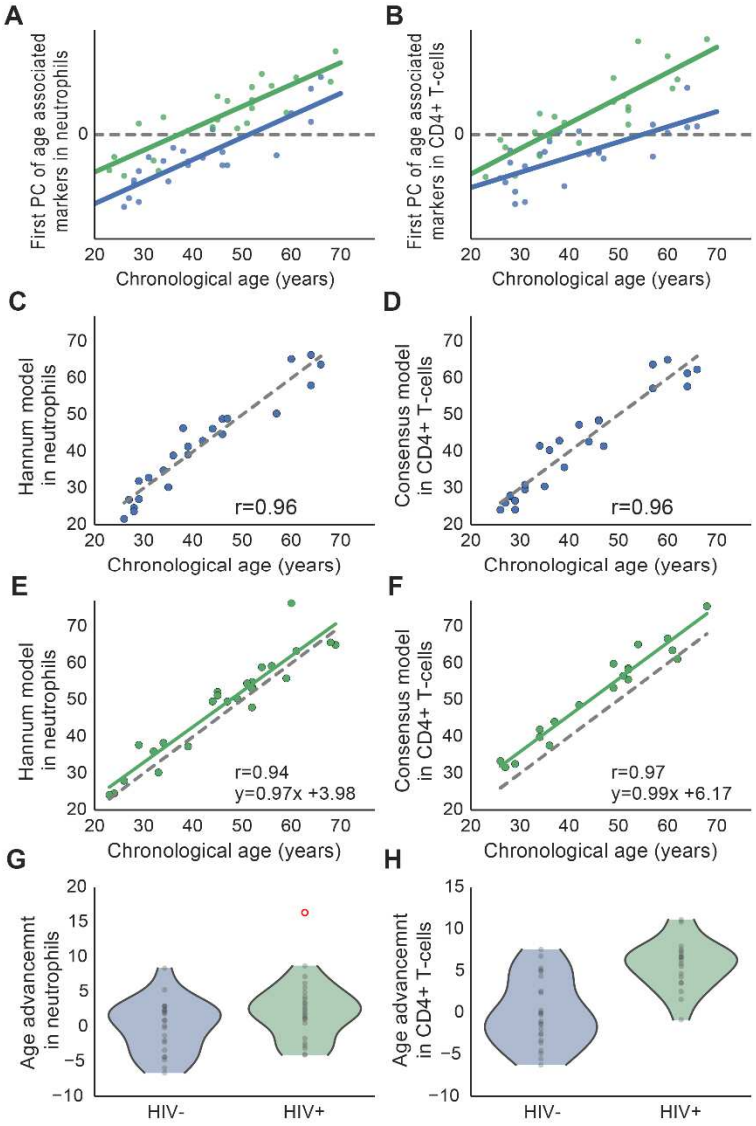


Figure 3.3: Age advancement in validation cohorts of purified cells. **A-B**, Unsupervised principal component (PC) analysis of methylation patterns in purified blood cell types, in which the first PC is positively associated with both age (x-axis) and disease status (HIV+, green; HIV- blue). (**A**) New CD4+ T-cell cohort across 5999 probes that are age-associated in $CD4^+$ T-cells (GSE59250). (**B**) New neutrophil cohort across 9971 probes that are age-associated in neutrophils (GSE65097). **C-F**, Control (**C-D**) and HIV+ (**E-F**) subjects for sorted cell validation datasets comparing chronological age to the Hannum et al. epigenetic aging model in neutrophils (**C,E**) and consensus aging model in $CD4^+$ T-cells (**D,F**). **G-H**, Violin plots showing age advancement in the two sorted cell datasets. For **B**, in initial analysis the first PC heavily reflected an outlier point, which was removed for this analysis after which the PC was recalculated.

Chapter 3.4.7: HIV and aging have shared and distinct methylation patterns

We identified 81,361 CpG markers associated with HIV infection (Benjamini-Hochberg corrected $P < 0.01$; likelihood ratio test using a multivariate linear model, **Supplementary Table 3.5**). Of these, 5631 were also associated with aging, a 1.3-fold enrichment over random expectation (**Figure 3.4A**, Fisher's Exact Test $P < 10^{-58}$, **Supplementary Table 3.6**). We found that markers associated with both HIV and aging were enriched in DNAse hypersensitivity sites and transcriptional start sites, suggesting methylation changes in DNA regions under active regulation. These CpG markers were also enriched in binding sites for polycomb repressive complex (PRC2) (**Figure 3.4B**), a switch that tightly regulates genes required for differentiation and renewal and in *Drosophilia* is linked to longevity (Siebold *et al.*, 2010). These findings reinforce previous reports that PRC2 targets are permanently repressed by methylation during the aging process (Beerman *et al.*, 2013; Deaton and Bird, 2011; Teschendorff *et al.*, 2010). Interestingly, markers associated with HIV but not aging had a very different functional enrichment profile (**Figure 3.4B**), indicating additional mechanism(s) for epigenetic alteration associated with HIV.

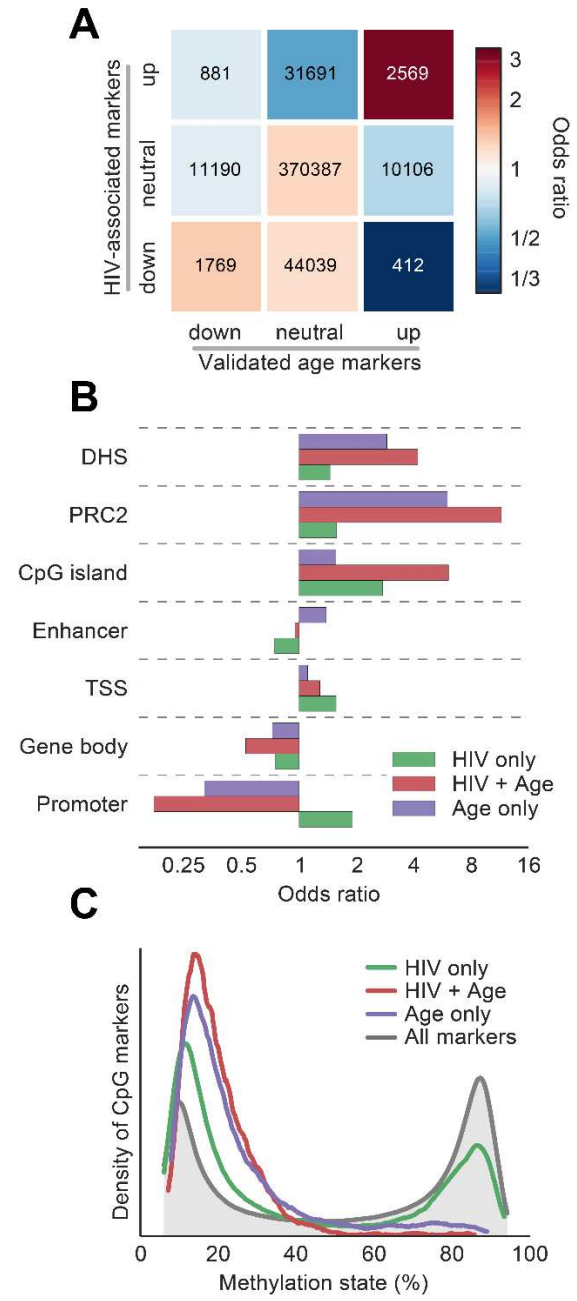


Figure 3.4: HIV and aging have shared and distinct methylation patterns. A, Overlap table comparing the set of CpG markers associated with HIV and the set of validated age-associated markers (see **Figure 3.1A**). Numbers indicate probe counts in each overlap, colors correspond to odds ratio of overlap compared to background. **B**, Odds ratios of enrichment for a panel of genomic features, evaluated in sets of markers associated with age, HIV, or both. PRC2, polycomb repressive complex 2 binding sites; DHS, DNase hypersensitivity sites; TSS, transcription start sites. **C**, Distribution of methylation states for the CpG marker sets defined in (**A**).

We have previously reported that age-associated markers in older subjects tend away from a fully methylated or unmethylated state and instead move towards disorder (50% methylation) (Hannum *et al.*, 2012). We found that HIV-infected patients displayed a similar trait: among markers associated with HIV, 66% tended towards disorder, compared with 70% of age-associated markers (**Experimental Procedures**, **Supplementary Figure 3.6**). Furthermore, whereas age-associated markers tended to have a low methylation fraction that increased with age, HIV-associated markers were more equally balanced between low and high methylation states (**Figure 3.4C**).

Chapter 3.4.8: HIV is associated with hypomethylation of the HLA locus

Taking into account this general increase in disorder, we sought to determine if there were any specific genomic regions for which the methylation state was particularly associated with HIV infection. An epigenome-wide screen of whole blood identified a single genomic region that was enriched in CpG markers associated with HIV; this region, consisting of 10 megabases on chromosome 6 including histone gene cluster 1 and the entire HLA locus, had particularly reduced methylation levels in HIV+ cases as compared to HIV- controls ($P < 10^{-10}$, **Figure 3.5A**, **Experimental Procedures**). HLA genes encode the Major Histocompatibility Complexes (MHC), the key antigen-presenting molecules that govern the acquired immune response and impact innate immunity (**Figure 3.5B**) (Goulder and Walker, 2012). We found that the differentially methylated markers surround the rs2395029 variant, for which common genetic variation has been repeatedly implicated in HIV host control (**Figures 3.5C,D)** (Fellay *et al.*, 2007;

International HIV Controllers Study *et al.*, 2010). Examination of this locus in the validation samples of purified neutrophils and $CD4^+$ T cells identified the HCP5 gene body as particularly differentially methylated (**Figures 3.5E,F, Supplementary Figure 3.7**). As further evidence that the observed changes are functional, we found that the amount of methylation at this gene was correlated with a patient's $CD4^+$ / $CD8^+$ T cell ratio (**Supplementary Figure 3.7**). An intriguing interpretation of our results is that some of the previously reported changes in HLA expression and corresponding HIV control (Apps *et al.*, 2013) are attributable to methylation dynamics.

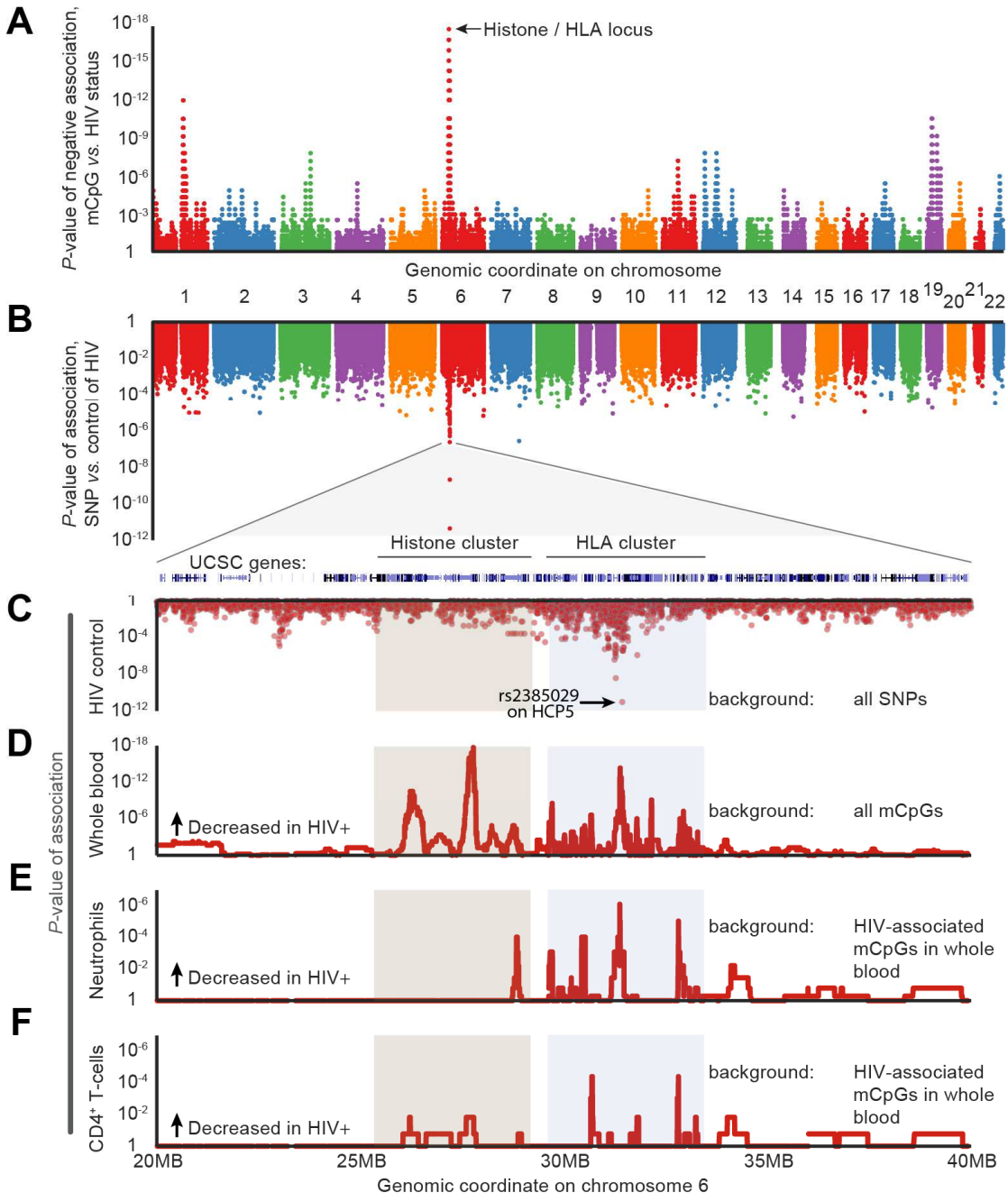


Figure 3.5: Methylome remodeling under sustained HIV infection targets HLA. A, Epigenome-wide association of CpG methylation (mCpG) with HIV status (presence or absence). Each point represents the P-value of enrichment for differentially methylated CpG markers within a bin of ±100 consecutive markers along the genome. **B-C**, P-values of genome-wide association of single nucleotide polymorphisms (SNP) with host control of HIV, reproduced from Fellay et al. **D**, Epigenome-wide association of mCpG with HIV status (presence or absence), zoomed in to target histone / HLA locus. **E-F**, Validation screen of HIV-downregulated markers in neutrophils (**E**) and CD4+ T-cells (**F**).

Chapter 3.5: Discussion

Here, we have shown that methylome-wide changes previously ascribed to aging are also induced by HIV (**Figures 3.1 and 3.3**). By using highly accurate, externally trained and validated models of biological aging, our study provides a robust estimate of a five-year age advancement in HIV/cART individuals (**Figure 3.2**). These results, in combination with the link between molecular age advancement and increased mortality risk (Marioni *et al.*, 2015), support the idea that chronic HIV infection is accompanied by a tangible gerontological phenotype. In addition to an aggregate estimate of HIV age advancement, the methylation aging model allows for patient-by-patient estimates. Patients deemed more likely to suffer from HIV-mediated aging effects might be placed on alterative schedules for preventative care, including early screening and further testing if warranted.

While epidemiological studies have attempted to measure age acceleration and increased mortality rates in HIV+ individuals (Appay and Rowland-Jones, 2002; Guaraldi *et al.*, 2011; Smith *et al.*, 2012), such measurements are made difficult by the myriad co-factors associated with HIV infection. For instance, low $CD4^+$ T cell counts, HCV infection, and drug usage are important factors of HIV infection that are also suspected to significantly affect mortality rates. Most previous studies have not attempted to control for these factors; in contrast, our study has focused specifically on healthy white male non-drug users. Nonetheless, our estimate of HIV age advancement of 4.9 years, calculated from a quantitative analysis of the methylome, falls within the range of the previous epidemiological studies. Further work will be needed to understand if the observed epigenetic age

advancement is generalizable to broader slices of the HIV+ population, i.e. patients with complex co-morbidities such as drug use or additional viral infections.

This study is based on the same epigenetic model of biological aging as many others, including recent reports associating epigenetic aging with Down's Syndrome (Horvath *et al.*, 2015), traumatic stress (Boks *et al.*, 2015), and even all-cause mortality (Marioni *et al.*, 2015). Here, we implement key data-processing and analysis steps to improve the application of these models, which should aid in future applications. By minimizing the effects of cell type composition, we find better calibration of our control samples (**Figure 3.2B-D and Supplementary Table 3.2**), and the model is less affected by confounding associations such as the changing blood composition that occurs in HIV+ individuals (**Supplementary Figure 3.3**). Furthermore, integration of both the Hannum *et al.* and Horvath (2013) models of epigenetic aging serves to limit biases in model training and allows us to filter samples that are of low quality or ill-suited for use in aging studies (**Experimental Procedures, Supplementary Table 3.2**).

Our finding of a five-year age advancement in cART-treated subjects (**Figure 3.2E**) is similar to one recent report (Horvath and Levine, 2015) but contrasts with another study in untreated patients, in which shared effects of age and HIV on the methylome were used to report an age advancement of 14 years (Rickabaugh *et al.*, 2014). Although this discrepancy could be due to a beneficial effect of cART, we believe it is more likely due to differing statistical approaches. The previous number is based on comparison of the effects of HIV and age in a single cohort, rather than an epigenetic model of aging built for normal individuals,

as performed here. Moreover, the authors derive their estimate from the ratio of linear coefficients for HIV and age, which are themselves highly correlated; such co-linearity is a well-known cause of instability in such estimates (Farrar and Glauber, 1967).

The discovery of HLA hypomethylation as a targeted consequence of HIV infection (**Figure 3.5**) has compelling synergy with the earlier discoveries of HLA genotype and expression level as major determinants of HIV control. Common genetic variation in HLA has been identified as the major contributing factor to host control of HIV infection (Fellay *et al.*, 2007; International HIV Controllers Study *et al.*, 2010), and HLA has been reported as a hotspot for integration of HIV provirus (Ambrosi *et al.*, 2011). HIV infection has also been associated with decreased expression of some HLA genes but not others (Bonaparte and Barker, 2004; Cohen *et al.*, 1999), and higher HLA-C expression is associated with HIV control (Apps *et al.*, 2013; Kulkarni *et al.*, 2011; Thomas *et al.*, 2009). Our result suggests an epigenetic component to the regulation of HLA expression in this region. It also raises the possibility that the ability to control HIV infection could be acquired through epigenetic modification, as well as inherited through genotype.

In summary, we have shown that an extrinsic perturbation to a human population, driven by HIV infection and cART, is capable of inducing changes in the epigenomic state of affected individuals. This perturbation may influence regulation of HLA gene expression and also encompasses signatures of aging. Our findings help address a long-standing debate regarding the effects of HIV infection on biological aging in cART-treated individuals, in a manner that can be

assessed numerically using an epigenome-based readout. Taken together, our findings show that the epigenome adds a quantitative means of assessing the interaction of HIV with normal and pathogenic processes associated with aging, and they shed light on the underlying mechanisms by which acute and chronic viral infection impact the host.

Chapter 3.6: Experimental procedures

Chapter 3.6.1: Reproduction of computation procedures

All data retrieval and processing steps are documented in a series of IPython notebooks at www.github.com/theandygross/HIV_Methylation. These notebooks provide fully executable instructions for the reproduction of the analyses and the generation of figures and statistics for this study.

Chapter 3.6.2: Selection criteria and subject recruitment

HIV+ subject samples were obtained from CHARTER as a Resource (www.charterresource.ucsd.edu). The CHARTER study was comprised of HIV-infected participants at varying stages of disease and with differing histories of antiretroviral treatment, with a focus on neuromedical and neurobehavioral assessments (Heaton *et al.*, 2010). We requested information on subjects for which DNA had been obtained. Demographic and clinical data were filtered for non-Hispanic white males (to match the control group) who were free of Hepatitis C virus, not diabetic, on cART, and adherent to therapy. These subjects had estimated time from HIV infection to sample collection of 0.2 - 26.1 years. Two

groups were selected for study, those more recently infected by HIV (but after the acute infection stage, 0.8 - 5.0 years of infection) and those chronically infected (>12.0 years). As a control, 44 non-Hispanic white males without HIV were recruited from the San Diego area. Clinical and demographic data are presented in **Supplementary Table 3.1**.

For validation samples 35 HIV+ subjects along with 25 healthy controls were recruited prospectively for the purpose of this study to match the characteristics of the primary cohort. Cells were purified using immunomagnetic separation and DNA was extracted from purified cell populations. While most subjects used had both neutrophils and $CD4^+$ T-cells profiled, differing DNA yield for some subjects prohibited profiling of both cell types for some patients.

Chapter 3.6.3: Sample collection and methylation analysis

DNA was purified from whole blood samples using PaxGene collection tubes (Qiagene) and FlexiGene DNA extraction kits (Qiagen). Methylation analysis was performed using Infinium HumanMethylation450 BeadChip Kits (Illumina). 500ng of DNA was bisulfite converted using EZ DNA Methylation Kits (Zymo Research) and subsequently processed for HumanMethylation450 BeadChips following manufacturer's instructions. Following hybridization, BeadChips were scanned using the Illumina HiScan System.

Chapter 3.6.4: Data pre-processing

All methylation data for HIV+ and HIV– subjects were deposited in the Gene Expression Omnibus (GEO) under accession number GSE67705. For the Hannum *et al.* (Hannum *et al.*, 2012) and EPIC (Riboli *et al.*, 2002) studies, raw data were

obtained from GEO accessions GSE40279 and GSE51032. All data were processed through the Minfi R processing pipeline (Aryee *et al.*, 2014). Cell counts were estimated by the estimateCellCounts function in Minfi using flow sorted cell populations made available by Houseman *et al.* (Houseman *et al.*, 2012). To limit variability in methylation levels due to differing cell type composition in the whole blood samples, methylation levels were adjusted for each CpG marker as follows:

- Average methylation levels for each cell type were obtained from the Houseman *et al.* (Houseman *et al.*, 2012) flow sorted blood dataset.
- A theoretical methylation level was assessed for each patient by assuming their blood to be a mixture of these pure cell populations at the estimated cell type proportions.
- The difference of each patient's methylation level from the average was assessed.
- This difference was subtracted from the original raw dataset.

We followed the protocol established to be optimal by Marabita *et al.* (Marabita *et al.*, 2013) first quantile normalizing the data and then performing beta-mixture quantile (BMIQ) normalization (Teschendorff *et al.*, 2013). To limit batch effects, all arrays across the three studies were normalized together. For use in the Horvath methylation age model, raw data were normalized to a gold standard reference distribution following the protocol provided in the manuscript (Horvath, 2013). The sole deviation from the Horvath protocol was an additional cell composition adjustment performed in a similar manner as described above, after BMIQ normalization. While the cell composition adjustment was not part of either

the Horvath or Hannum *et al.* processing pipeline, recent work (Jaffe and Irizarry, 2014) has shown cell type composition to be a key confounding factor in methylation analysis.

Chapter 3.6.5: Benchmarking the aging models

Aging models were assessed using the Hannum *et al.* and EPIC (Riboli *et al.*, 2002) datasets. While the Hannum *et al.* dataset was used to train both epigenetic aging models, the EPIC data were made available after the time of construction of both models and thus provide an independent assessment of performance. We limited analysis to patients between the ages of 25 and 68 years of age for better comparison to the HIV cohort. Among the HIV and EPIC cohorts, we saw slightly better performance of the Hannum model (which was trained using only whole blood data) than the Horvath model (trained in a variety of tissues), but when a simple average of these two models was taken (the 'consensus model'), we found better performance than either separately (**Supplementary Table 3.2**).

Chapter 3.6.6: Epigenetic model concordance filter

One key drawback of current models of molecular age is the lack of a confidence measure in model prediction for any particular individual. In the application of such models, it is often desirable to understand which predictions are of poor quality and should be treated with skepticism. To address, this we utilized the concordance between the two models as an additional filter of data quality. Despite general agreement between the models, we found a number of subjects for which the biological age predictions varied by more than 20%. We suspect that these samples were of poor data quality, or that the patients had

molecular effects of aging that were not properly trained into one or both of the methylation age models. This analysis resulted in the filtering of three HIV+ cases and two HIV− controls in our primary cohort (**Supplementary Table 3.1**).

Chapter 3.6.7: Linear scaling of epigenetic age

For both models and across all datasets a linear scaling factor existed when comparing chronological verses biological age. In order to properly compare the performance of the models and to best calibrate them to our dataset, we performed a linear adjustment to all model fits for the control data to a unit slope with a zero intercept. Note that this affected the model error when compared in an absolute sense, but did not affect the correlation between biological and chronological age. In HIV+ patients, we adjusted this particular cohort to the regression fit of the matched controls. This operation allowed us to quantify the age advancement of the HIV+ patients within the context of the HIV− control samples that were processed with this cohort.

Chapter 3.6.8: Screening for differentially methylated markers in response to HIV infection

For the results described in **Figures 3.3 and 3.4,** we ran a multivariate linear model to test for differentially methylated markers in response to HIV infection. This model used predicted cell type composition and age as covariates. Significance was assessed via a likelihood-ratio test for the improvement of a model fit with HIV as the variable of interest.

Chapter 3.6.9: Disorder of Methylation in Response to HIV and Aging

In addition to an age advancement phenotype in HIV+ patients (**Figures 3.1 and 3.2**), we observed increasing disorder of the methylome by both aging and HIV infection (**Supplementary Figure 3.4**, **Main Text**). Thus, we considered that the increasing age advancement might be explained by increasing disorder. To assess this possibility, we conducted a principal component analysis on 7967 age-associated markers that trended away from disorder with age (i.e., decreasing methylation for markers with fractions < 50%, increasing methylation for markers with fractions > 50%, **Supplementary Figure 3.4A**). This analysis produced a similar result to that shown in **Figure 3.1C**, in which the first principal component of the cohort was still associated with both with age and HIV infection.

Another way to investigate the potential effects of entropy on the aging phenotype is to examine the markers used by the biological age models. We observed that only 231 of 436 (53%) markers used in the two aging models tended towards methylation values of 50% (i.e. disorder, binomial P = 0.2). Furthermore, we calculated the Shannon entropy for each HIV+ and HIV– subject across the full set of CpG markers, as well as only the 436 markers used in the biological age models. From this analysis it can be seen that while there is a general entropy increase in HIV+ patients across the entire methylome (**Supplementary Figure 3.4B**), there is no such effect in the markers used in the biological age models (**Supplementary Figure 3.4C**). While increasing methylome disorder has a clear association with age, it seems the model selection procedure employed by both Hannum *et al.* (Hannum *et al.*, 2012) and Horvath (Horvath, 2013) used this feature

sparingly. From these analyses we can conclude that the increase in epigenetic age among HIV+ patients is a distinct biological signal, as opposed to a consequence of a shared molecular phenotype between HIV and aging.

Chapter 3.6.10: Identification of differentially methylated regions

In **Figure 3.4A**, we aimed to find differentially methylated regions associated with HIV infection. As described above and in the main text, we found that the majority of HIV-associated markers became more disordered (trended towards 50% methylation) as compared to their values in HIV– control patients. While global deregulation is only one reason for methylation at a DNA site to trend in this direction, we considered it more likely that features that trend away from disorder are specific to HIV infection and not cellular stress as a whole. For this reason, we explicitly looked for genomic regions which trended away from disorder in our genome-wide screen. This analysis identified 25,491 markers that were associated with HIV, trended away from disorder, and were not associated with age.

For the discovery and visualization of differentially methylated regions (**Figure 3.4A**), we calculated a rolling statistic on the density of 'hits' in 200 marker windows. From this analysis it was clear that a genomic region encompassing the HLA and histone gene clusters was enriched for markers in our query set, and post-hoc analysis confirmed a strong enrichment in the genomic interval traditionally assigned to the HLA region (~29MB - 33MB on chromosome 6, odds ratio = 1.3, $P < 10^{-10}$). For further refinement and visualization of this signal we conducted a similar scan statistic on a section of chromosome 6 in **Figure 3.4D**.

In this targeted analysis we relaxed our criteria and looked for regions of consistent increases or decreases in methylation in HIV+ verses HIV– subjects. This analysis showed a number of ‘peaks’ of hypomethylation both in the histone gene region as well as near the HLA genes.

Chapter 3.6.11: Accounting for the probe density of the HLA region

One potential confounding factor of this analysis is the high density of markers in the HLA region due to the design of the Illumina chip. Taking the non-uninform density of the chip into account, the scan statistic searched for regions across a fixed number of markers as opposed to a fixed-width genomic interval. Despite this, it is possible that the tight clustering of markers in this region gave us more power to detect short differentially methylated regions within this genetic locus. The presence of two peaks in the histone cluster region directly upstream the HLA locus gives strong support to this being a specific effect. The density of probes in the histone region was typical compared to the rest of the genome, and the coincidence of these two signals being close to each other solely by chance is minimal.

Chapter 3.7: Author Contributions

HF and TI conceived of the study. HF, BMM and KZ organized the HIV+ and control patient cohorts. DC, KF, KL, KLJ, and MK performed the sample preparation and methylation profiling experiments. AMG, PAJ, HS, and TI

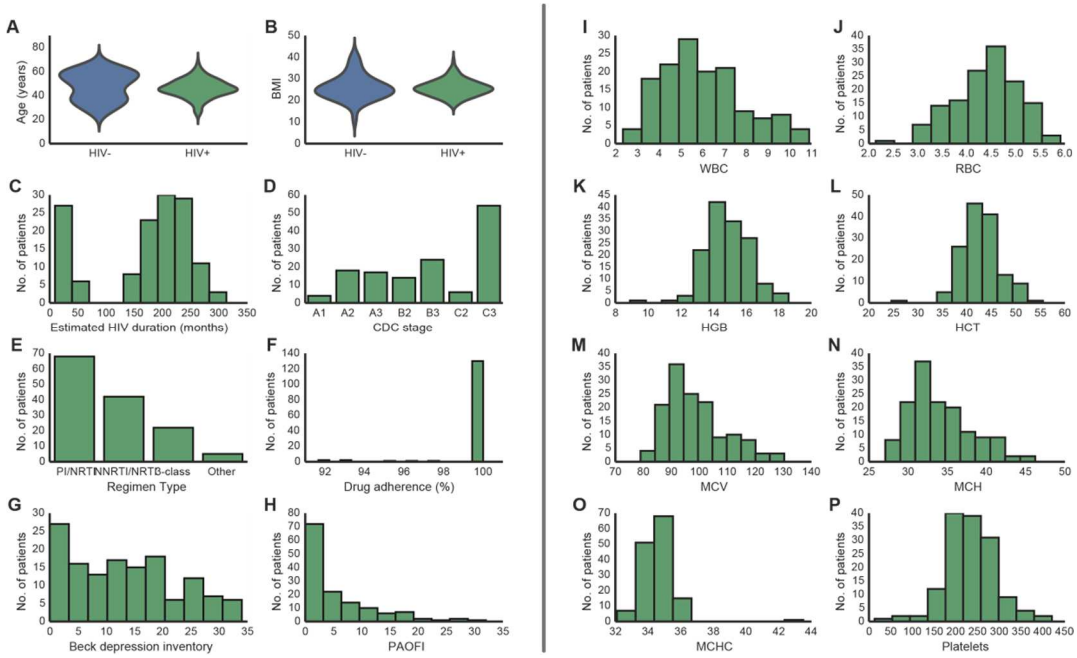
conducted study and analysis design. AMG performed all statistical and bioinformatics analyses. AMG, PAJ, JFK, HF and TI wrote the manuscript.

Chapter 3.8: Acknowledgements

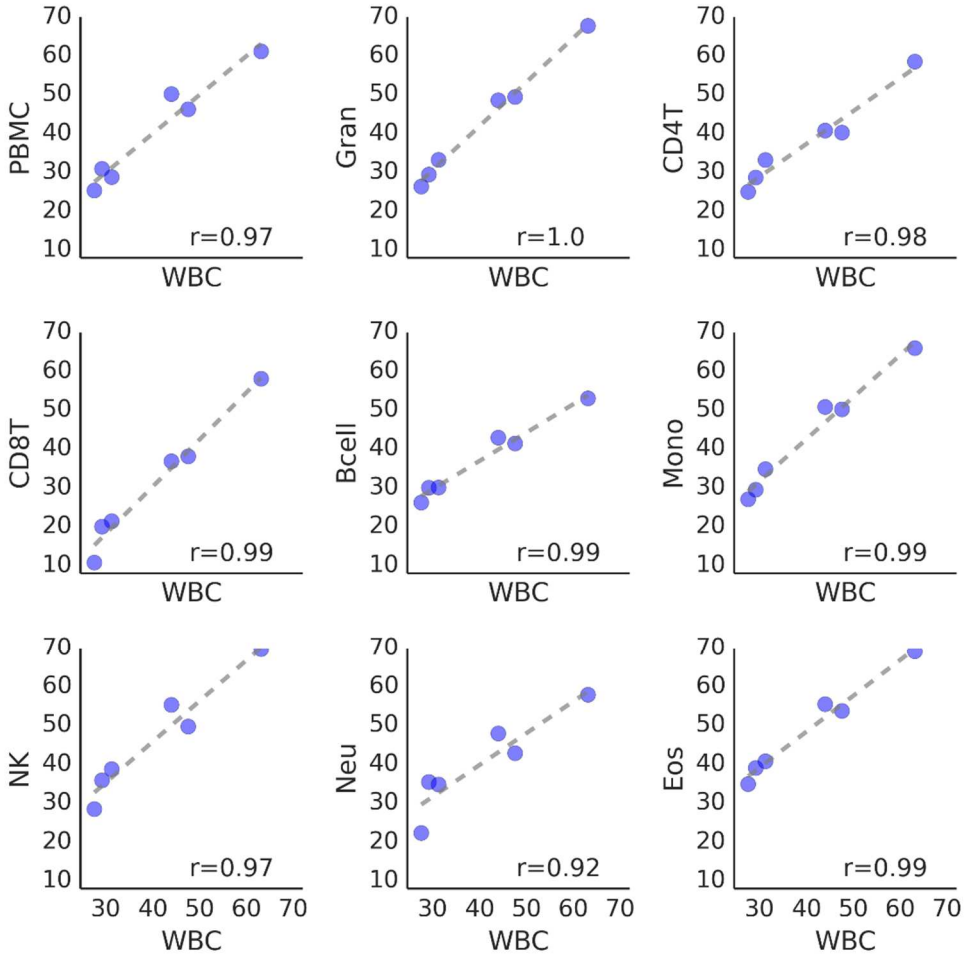
This work was supported by the National Institute of Mental Health (P30 MH062261), the National Cancer Institute (P30 CA023100) and the California Institute for Regenerative Medicine. The CHARTER study is supported by the National Institutes of Health (HHSN271201000030C). We wish to thank Roman Sasik and Aaron Chang for advice on methylome analysis, and Cherie Ng for constructive discussions and comments on the manuscript.

Chapter 3, in full, has been submitted for publication of the material as it may appear in Molecular Cell, 2015. Andrew M. Gross, Philipp A. Jaeger, Jason F. Kreisberg, Katherine Licon, Kristen L. Jepsen, Mahdieh Khosroheidari, Brenda M. Morsey, Hui Shen, Ken Flagg, Daniel Chen, Kang Zhang, Howard S. Fox, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

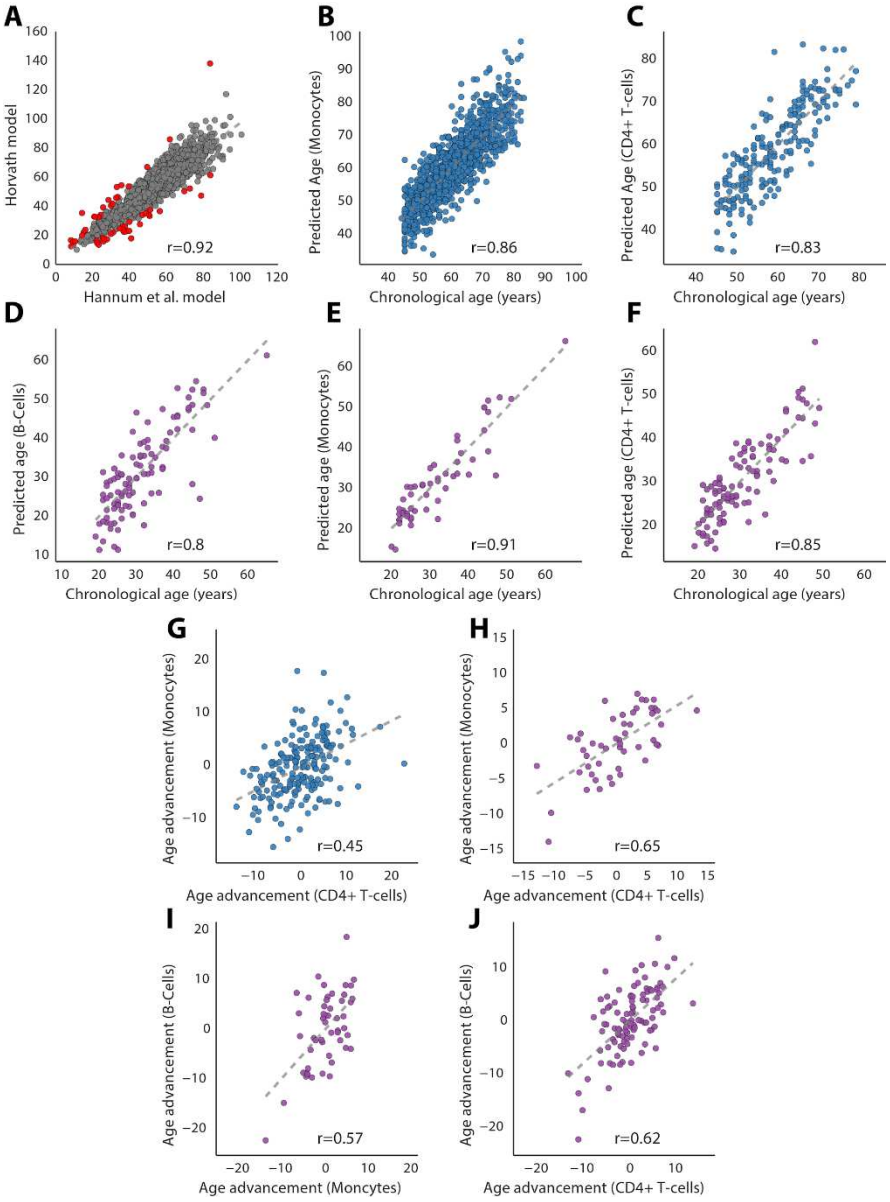
Chapter 3.9: Supplementary Figures



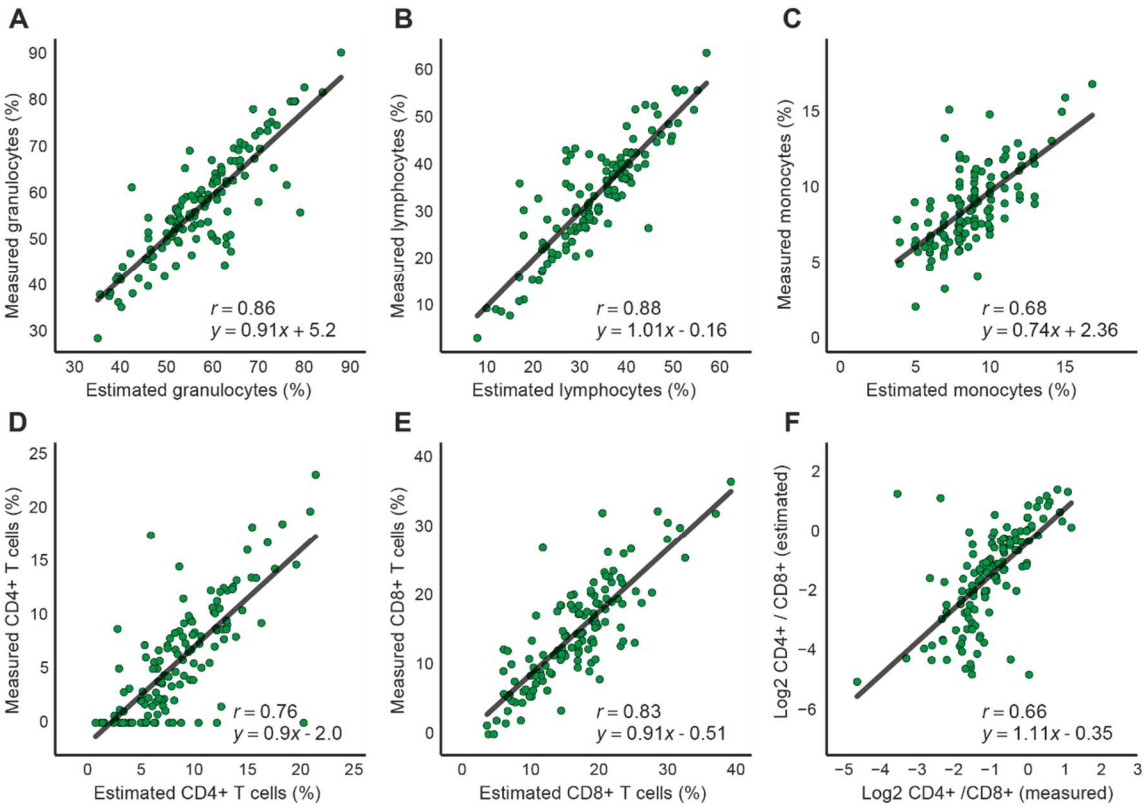
Supplementary Figure 3.1: Clinical variables and blood-based biomarkers collected for the primary cohort. A-B, Comparison of age and body mass index (BMI) across 137 cases and 42 control patients. **C-H,** Clinical variables collected from HIV-infected patients at the time of sample collection. HIV staging was based on standards of the Centers for Disease Control (CDC). Regime types indicate 2-class (PI, protease inhibitor; NRTI, nucleoside reverse-transcriptase inhibitor; NNRTI, non-nucleoside reverse-transcriptase inhibitor); 3-class, (PI/NRTI/NNRTI); PAOFI, patient's assessment of own functioning inventory. **I-P,** Complete blood counts from HIV-infected patients. WBC, white blood cell count; RBC, red blood cell count; HGB, hemoglobin; HCT, hematocrit; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration. For further characterization of HIV patients see **Supplementary Table 3.1**.



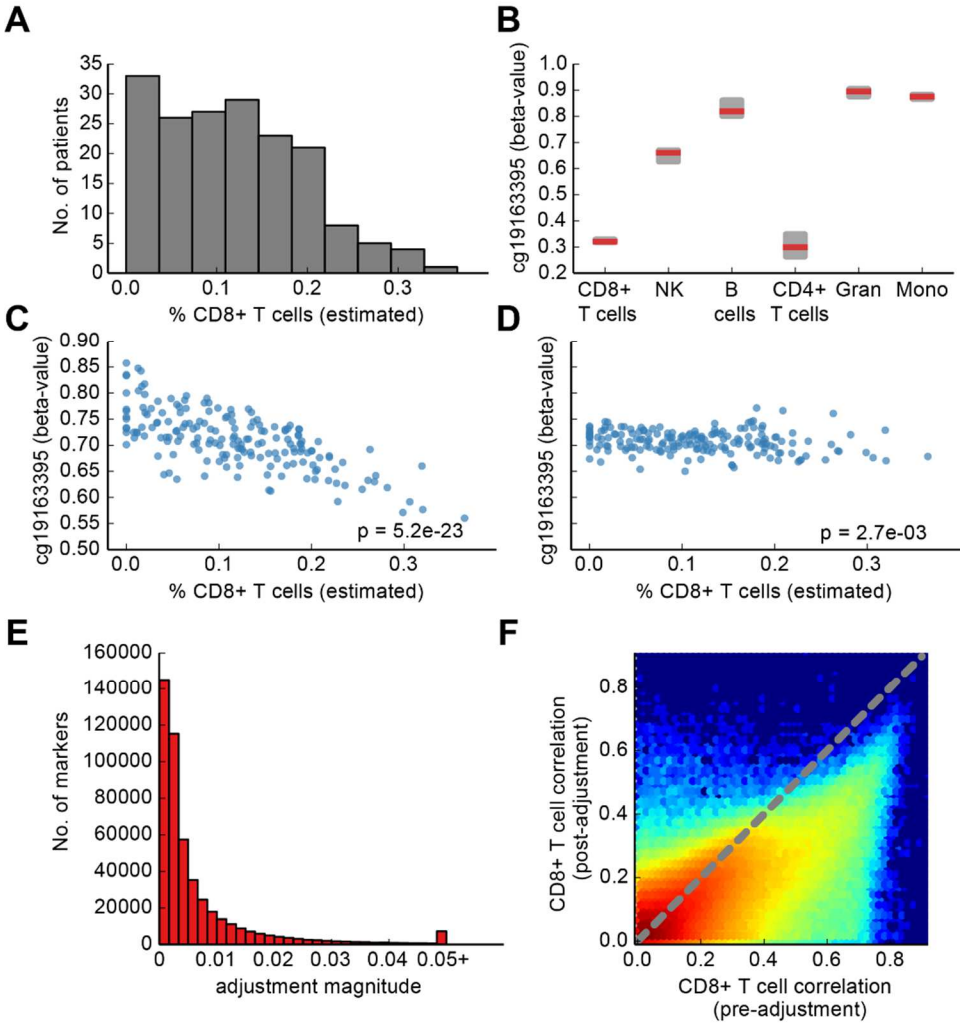
Supplementary Figure 3.2: Comparison of biological age calculated from whole blood versus purified cell types in six individuals. On the x-axis are the age predictions taken from whole-blood measurements; on the y-axis are age predictions in sorted cell populations for the same subjects. Methylation data taken from (Houseman et al., 2012). WBC, whole blood cells; PBMC, peripheral blood mononuclear cells; Gran, granulocytes; CD4T, CD4+ T cells; CD8T, CD8+ T cells; Bcell, B cells; Mono, monocytes; NK, natural killer cells; Neu, neutrophils; Eos, eosinophils.



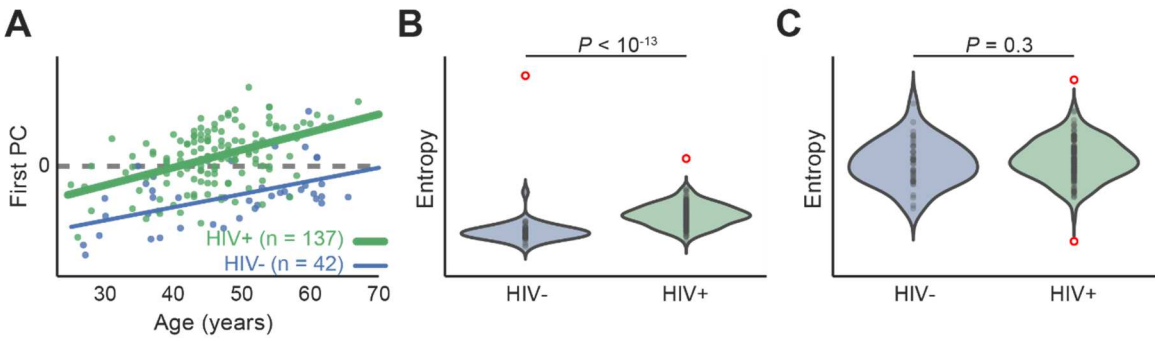
Supplementary Figure 3.3: Evaluation of epigenetic age predictions in sorted cell datasets. **A,** Scatter plot comparing the ages predicted using the Hannum *et al*. and Horvath models on 1688 samples obtained from sorted cell datasets. Red points indicate patients that were discarded due to disagreement between the two aging models (n=54). **B-C** Accuracy of the consensus model (y-axis) to predict true chronological age (x-axis) in sorted cell datasets from Reynolds *et al.* (*n*=1130 and 201 for monocytes and CD4+ T-cells, respectively). **D-F,** Accuracy of the consensus model (y-axis) to predict true chronological age (x-axis) in sorted cell datasets from Absher *et al.* (*n*=54, 103, and 102 for monocytes, B cells, and CD4+ T-cells, respectively). **G-H,** Scatter plots comparing age advancement in patient matched samples for Reynolds *et al.* (**G**) and Absher *et al.* (**H-J**).



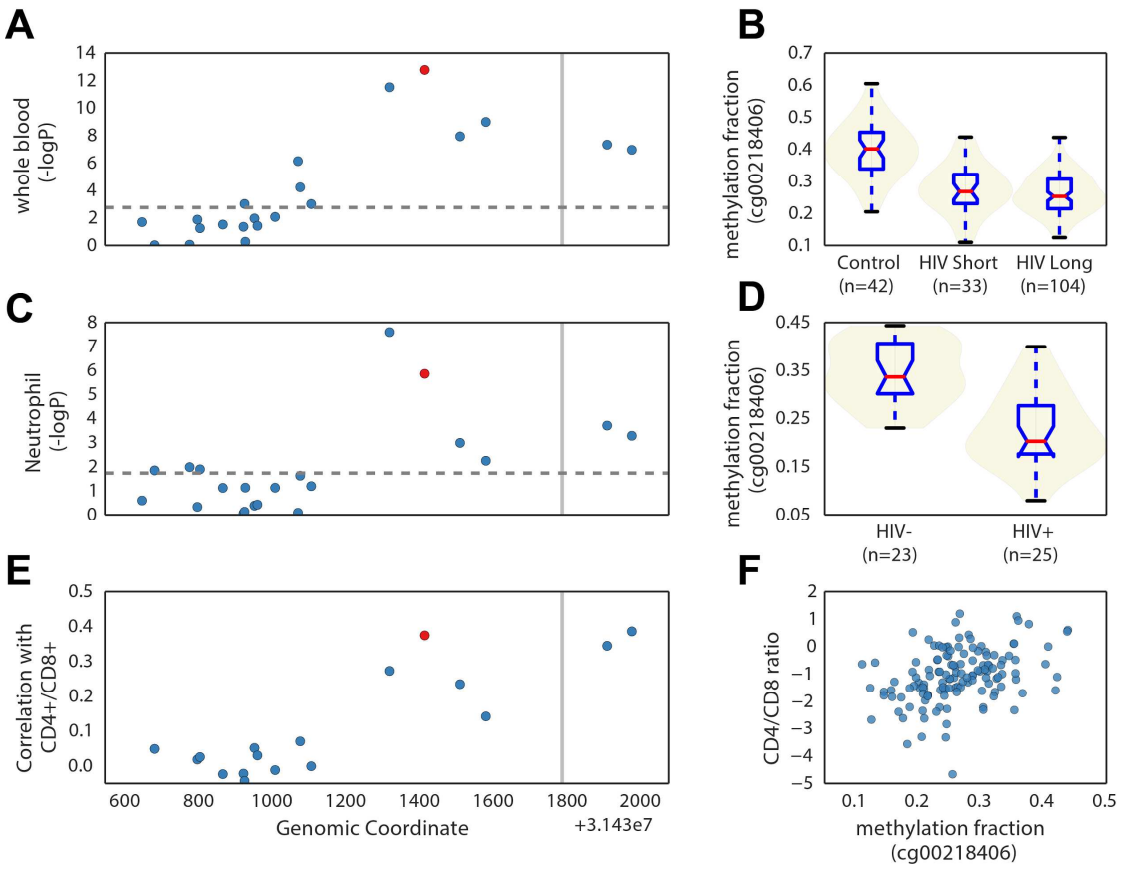
Supplementary Figure 3.4: Concordance of estimated cell counts with lab measured values in HIV-infected patients. Percentages of each cell type (**A**-**F**) were estimated using the Jaffe and Irizarry method. Black lines indicate regression.



Supplementary Figure 3.5: Summary of cell composition adjustment procedure. A, Distribution of estimated CD8+ T cell composition across the primary patient cohort (Methods). **B,** Distribution of CpG methylation fraction for marker cg19163395 across flow sorted blood populations for six individuals obtained from the Houseman *et al.* dataset. NK, Natural Killer cells; Gran, Granulocytes; Mono, Monocytes. **C-D,** Association of cg19163395 methylation fraction *versus* CD8+ T cell abundance in unadjusted (**C**) and adjusted (**D**) data. **E,** Distribution of the standard deviation of magnitude adjustment across the primary cohort for each marker measured. **F,** Hex-bin plot showing correlation of marker methylation levels (logit-adjusted beta values) with CD8+ T cell abundance before and after cell composition adjustment.



Supplementary Figure 3.6: Observed disorder (entropy) in different sets of CpG markers. A, First principal component (PC) of 7967 age associated markers that trend away from disorder across HIV-infected patients (green) and healthy controls (blue, **Methods**). **B,** Relative entropy comparing HIV+ to HIV– individuals across the 473,044 markers passing quality control. **C,** Relative entropy comparing HIV+ to HIV– individuals across 436 markers used at least one of the epigenetic models of aging. Significance assessed by Mann-Whitney U test. Note that red circles indicate outliers ±3 standard deviations away from the mean. These are not used to fit the violin profile, but are used in the statistical assessment.



Supplementary Figure 3.7: Exploration of methylation markers annotated to HCP5. A, Log p-values for association of methylation with HIV infection in primary whole blood samples. **B,** Violin plots showing the distribution of methylation values for cg0028406, the most HIV-associated marker in this region. **C,** Log p-values for association of methylation with HIV infection in the purified neutrophil samples. **D,** Violin plots showing the distribution of methylation values for cg0028406. **E,** Correlation of methylation values with CD4+/CD8+ T-cell ratio among HIV+ subjects. **F,** Scatter plot of CD4+/CD8+ T-cell ratio versus cg0028406 methylation levels. Vertical line corresponds to rs2395029, a SNP having major association with HIV host control. Red points in left-hand plots represent cg0028406, the probe profiled in the right-hand plots.

Chapter 3.10: References

Absher, D.M., Li, X., Waite, L.L., Gibson, A., Roberts, K., Edberg, J., Chatham, W.W., and Kimberly, R.P. (2013). Genome-Wide DNA Methylation Analysis of Systemic Lupus Erythematosus Reveals Persistent Hypomethylation of Interferon Genes and Compositional Changes to CD4+ T-cell Populations. PLoS Genetics 9, e1003678.

Althoff, K.N., McGinnis, K.A., Wyatt, C.M., Freiberg, M.S., Gilbert, C., Oursler, K.K., Rimland, D., Rodriguez-Barradas, M.C., Dubrow, R., Park, L.S., Skanderson,

M., Shiels, M.S., Gange, S.J., Gebo, K.A., Justice, A.C., and Veterans Aging Cohort Study (VACS) (2015). Comparison of risk and age at diagnosis of myocardial infarction, end-stage renal disease, and non-AIDS-defining cancer in HIV-infected versus uninfected adults. Clin. Infect. Dis. 60, 627–638.

Ambrosi, A., Glad, I.K., Pellin, D., Cattoglio, C., Mavilio, F., Di Serio, C., and Frigessi, A. (2011). Estimated Comparative Integration Hotspots Identify Different Behaviors of Retroviral Gene Transfer Vectors. PLoS Computational Biology 7, e1002292.

Appay, V., and Rowland-Jones, S.L. (2002). Premature ageing of the immune system: the cause of AIDS? Trends Immunol. 23, 580–585.

Apps, R., Qi, Y., Carlson, J.M., Chen, H., Gao, X., Thomas, R., Yuki, Y., Del Prete, G.Q., Goulder, P., Brumme, Z.L., Brumme, C.J., John, M., Mallal, S., Nelson, G., Bosch, R., Heckerman, D., Stein, J.L., Soderberg, K.A., Moody, M.A., Denny, T.N., Zeng, X., Fang, J., Moffett, A., Lifson, J.D., Goedert, J.J., Buchbinder, S., Kirk, G.D., Fellay, J., McLaren, P., Deeks, S.G., Pereyra, F., Walker, B., Michael, N.L., Weintrob, A., Wolinsky, S., Liao, W., and Carrington, M. (2013). Influence of HLA-C expression level on HIV control. Science 340, 87–91.

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30, 1363–1369.

Barouch, D.H., and Deeks, S.G. (2014). Immunologic strategies for HIV-1 remission and eradication. Science 345, 169–174.

Beerman, I., Bock, C., Garrison, B.S., Smith, Z.D., Gu, H., Meissner, A., and Rossi, D.J. (2013). Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. Cell Stem Cell 12, 413–425.

Boks, M.P., van Mierlo, H.C., Rutten, B.P.F., Radstake, T.R.D.J., De Witte, L., Geuze, E., Horvath, S., Schalkwyk, L.C., Vinkers, C.H., Broen, J.C.A., and Vermetten, E. (2015). Longitudinal changes of telomere length and epigenetic age related to traumatic stress and post-traumatic stress disorder. Psychoneuroendocrinology 51, 506–512.

Bonaparte, M.I., and Barker, E. (2004). Killing of human immunodeficiency virus-infected primary T-cell blasts by autologous natural killer cells is dependent on the ability of the virus to alter the expression of major histocompatibility complex class I molecules. Blood 104, 2087–2094.

Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J., Lieberman, J., and Elledge, S.J. (2008). Identification of Host Proteins Required for HIV Infection Through a Functional Genomic Screen. Science 319, 921–926.

Bushman, F.D., Malani, N., Fernandes, J., D'Orso, I., Cagney, G., Diamond, T.L., Zhou, H., Hazuda, D.J., Espeseth, A.S., König, R., Bandyopadhyay, S., Ideker, T., Goff, S.P., Krogan, N.J., Frankel, A.D., Young, J.A.T., and Chanda, S.K. (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. PLoS Pathog. 5, e1000437.

Cao, W., Jamieson, B.D., Hultin, L.E., Hultin, P.M., Effros, R.B., and Detels, R. (2009). Premature Aging of T cells Is Associated With Faster HIV-1 Disease Progression: JAIDS Journal of Acquired Immune Deficiency Syndromes 50, 137–147.

Cawthon, R.M., Smith, K.R., O'Brien, E., Sivatchenko, A., and Kerber, R.A. (2003). Association between telomere length in blood and mortality in people aged 60 years or older. The Lancet 361, 393–395.

Chou, J.P., Ramirez, C.M., Wu, J.E., and Effros, R.B. (2013). Accelerated aging in HIV/AIDS: novel biomarkers of senescent human CD8+ T cells. PLoS ONE 8, e64702.

Christensen, B.C., Houseman, E.A., Marsit, C.J., Zheng, S., Wrensch, M.R., Wiemels, J.L., Nelson, H.H., Karagas, M.R., Padbury, J.F., Bueno, R., Sugarbaker, D.J., Yeh, R.-F., Wiencke, J.K., and Kelsey, K.T. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 5, e1000602.

Cohen, G.B., Gandhi, R.T., Davis, D.M., Mandelboim, O., Chen, B.K., Strominger, J.L., and Baltimore, D. (1999). The selective downregulation of class I major histocompatibility complex proteins by HIV-1 protects HIV-infected cells from NK cells. Immunity 10, 661–671.

Cuzin, L., Delpierre, C., Gerard, S., Massip, P., and Marchou, B. (2007). Immunologic and Clinical Responses to Highly Active Antiretroviral Therapy in Patients with HIV Infection Aged >50 Years. Clinical Infectious Diseases 45, 654–657.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes & Development 25, 1010–1022.

Deeks, S.G. (2011). HIV infection, inflammation, immunosenescence, and aging. Annu. Rev. Med. 62, 141–155.

Deeks, S.G., Lewin, S.R., and Havlir, D.V. (2013). The end of AIDS: HIV infection as a chronic disease. The Lancet 382, 1525–1533.

Deeks, S.G., Tracy, R., and Douek, D.C. (2013). Systemic effects of inflammation on health during chronic HIV infection. Immunity 39, 633–645.

Dubrow, R., Silverberg, M.J., Park, L.S., Crothers, K., and Justice, A.C. (2012). HIV infection, aging, and immune function: implications for cancer risk and prevention. Curr Opin Oncol 24, 506–516.

Emig-Agius, D., Olivieri, K., Pache, L., Shih, H.L., Pustovalova, O., Bessarabova, M., Young, J.A.T., Chanda, S.K., and Ideker, T. (2014). An integrated map of HIV-human protein complexes that facilitate viral infection. PLoS ONE 9, e96687.

Farrar, D.E., and Glauber, R.R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. The Review of Economics and Statistics 49, 92–107.

Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Francioli, P., Mallal, S., Martinez-Picado, J., Miro, J.M., Obel, N., Smith, J.P., Wyniger, J., Descombes, P., Antonarakis, S.E., Letvin, N.L., McMichael, A.J., Haynes, B.F., Telenti, A., and Goldstein, D.B. (2007). A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. Science 317, 944–947.

Freiberg, M.S., Chang, C.-C.H., Kuller, L.H., Skanderson, M., Lowy, E., Kraemer, K.L., Butt, A.A., Bidwell Goetz, M., Leaf, D., Oursler, K.A., Rimland, D., Rodriguez Barradas, M., Brown, S., Gibert, C., McGinnis, K., Crothers, K., Sico, J., Crane, H., Warner, A., Gottlieb, S., Gottdiener, J., Tracy, R.P., Budoff, M., Watson, C., Armah, K.A., Doebler, D., Bryant, K., and Justice, A.C. (2013). HIV infection and the risk of acute myocardial infarction. JAMA Intern Med 173, 614–622.

Goulder, P.J.R., and Walker, B.D. (2012). HIV and HLA class I: an evolving relationship. Immunity 37, 426–440.

Guaraldi, G., Orlando, G., Zona, S., Menozzi, M., Carli, F., Garlassi, E., Berti, A., Rossi, E., Roverato, A., and Palella, F. (2011). Premature age-related comorbidities among HIV-infected persons compared with the general population. Clin. Infect. Dis. 53, 1120–1126.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., and Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell 49, 359–367.

Heaton, R.K., Clifford, D.B., Franklin, D.R., Woods, S.P., Ake, C., Vaida, F., Ellis, R.J., Letendre, S.L., Marcotte, T.D., Atkinson, J.H., Rivera-Mindt, M., Vigil, O.R., Taylor, M.J., Collier, A.C., Marra, C.M., Gelman, B.B., McArthur, J.C., Morgello, S., Simpson, D.M., McCutchan, J.A., Abramson, I., Gamst, A., Fennema-

Notestine, C., Jernigan, T.L., Wong, J., Grant, I., and CHARTER Group (2010). HIV-associated neurocognitive disorders persist in the era of potent antiretroviral therapy: CHARTER Study. Neurology 75, 2087–2096.

Heyn, H., Li, N., Ferreira, H.J., Moran, S., Pisano, D.G., Gomez, A., Diez, J., Sanchez-Mut, J.V., Setien, F., Carmona, F.J., Puca, A.A., Sayols, S., Pujana, M.A., Serra-Musach, J., Iglesias-Platas, I., Formiga, F., Fernandez, A.F., Fraga, M.F., Heath, S.C., Valencia, A., Gut, I.G., Wang, J., and Esteller, M. (2012). Distinct DNA methylomes of newborns and centenarians. Proceedings of the National Academy of Sciences 109, 10522–10527.

Hilton, R. (2013). Human immunodeficiency virus infection and kidney disease. The Journal of the Royal College of Physicians Od Edinburgh 43, 236–240.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biology 14, R115.

Horvath, S., Erhart, W., Brosch, M., Ammerpohl, O., von Schonfels, W., Ahrens, M., Heits, N., Bell, J.T., Tsai, P.-C., Spector, T.D., Deloukas, P., Siebert, R., Sipos, B., Becker, T., Rocken, C., Schafmayer, C., and Hampe, J. (2014). Obesity accelerates epigenetic aging of human liver. Proceedings of the National Academy of Sciences 111, 15538–15543.

Horvath, S., Garagnani, P., Bacalini, M.G., Pirazzini, C., Salvioli, S., Gentilini, D., Di Blasio, A.M., Giuliani, C., Tung, S., Vinters, H.V., and Franceschi, C. (2015). Accelerated epigenetic aging in Down syndrome. Aging Cell n/a – n/a.

Houseman, E., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13, 86.

International HIV Controllers Study, Pereyra, F., Jia, X., McLaren, P.J., Telenti, A., de Bakker, P.I.W., Walker, B.D., Ripke, S., Brumme, C.J., Pulit, S.L., Carrington, M., Kadie, C.M., Carlson, J.M., Heckerman, D., Graham, R.R., Plenge, R.M., Deeks, S.G., Gianniny, L., Crawford, G., Sullivan, J., Gonzalez, E., Davies, L., Camargo, A., Moore, J.M., Beattie, N., Gupta, S., Crenshaw, A., Burtt, N.P., Guiducci, C., Gupta, N., Gao, X., Qi, Y., Yuki, Y., Piechocka-Trocha, A., Cutrell, E., Rosenberg, R., Moss, K.L., Lemay, P., O'Leary, J., Schaefer, T., Verma, P., Toth, I., Block, B., Baker, B., Rothchild, A., Lian, J., Proudfoot, J., Alvino, D.M.L., Vine, S., Addo, M.M., Allen, T.M., Altfeld, M., Henn, M.R., Le Gall, S., Streeck, H., Haas, D.W., Kuritzkes, D.R., Robbins, G.K., Shafer, R.W., Gulick, R.M., Shikuma, C.M., Haubrich, R., Riddler, S., Sax, P.E., Daar, E.S., Ribaudo, H.J., Agan, B., Agarwal, S., Ahern, R.L., Allen, B.L., Altidor, S., Altschuler, E.L., Ambardar, S., Anastos, K., Anderson, B., Anderson, V., Andrady, U., Antoniskis, D., Bangsberg, D., Barbaro, D., Barrie, W., Bartczak, J., Barton, S., Basden, P., Basgoz, N.,

Bazner, S., Bellos, N.C., Benson, A.M., Berger, J., Bernard, N.F., Bernard, A.M., Birch, C., Bodner, S.J., Bolan, R.K., Boudreaux, E.T., Bradley, M., Braun, J.F., Brndjar, J.E., Brown, S.J., Brown, K., Brown, S.T., Burack, J., Bush, L.M., Cafaro, V., Campbell, O., Campbell, J., Carlson, R.H., Carmichael, J.K., Casey, K.K., Cavacuiti, C., Celestin, G., Chambers, S.T., Chez, N., Chirch, L.M., Cimoch, P.J., Cohen, D., Cohn, L.E., Conway, B., Cooper, D.A., Cornelson, B., Cox, D.T., Cristofano, M.V., Cuchural, G., Czartoski, J.L., Dahman, J.M., Daly, J.S., Davis, B.T., Davis, K., Davod, S.M., DeJesus, E., Dietz, C.A., Dunham, E., Dunn, M.E., Ellerin, T.B., Eron, J.J., Fangman, J.J.W., Farel, C.E., Ferlazzo, H., Fidler, S., Fleenor-Ford, A., Frankel, R., Freedberg, K.A., French, N.K., Fuchs, J.D., Fuller, J.D., Gaberman, J., Gallant, J.E., Gandhi, R.T., Garcia, E., Garmon, D., Gathe, J.C., Gaultier, C.R., Gebre, W., Gilman, F.D., Gilson, I., Goepfert, P.A., Gottlieb, M.S., Goulston, C., Groger, R.K., Gurley, T.D., Haber, S., Hardwicke, R., Hardy, W.D., Harrigan, P.R., Hawkins, T.N., Heath, S., Hecht, F.M., Henry, W.K., Hladek, M., Hoffman, R.P., Horton, J.M., Hsu, R.K., Huhn, G.D., Hunt, P., Hupert, M.J., Illeman, M.L., Jaeger, H., Jellinger, R.M., John, M., Johnson, J.A., Johnson, K.L., Johnson, H., Johnson, K., Joly, J., Jordan, W.C., Kauffman, C.A., Khanlou, H., Killian, R.K., Kim, A.Y., Kim, D.D., Kinder, C.A., Kirchner, J.T., Kogelman, L., Kojic, E.M., Korthuis, P.T., Kurisu, W., Kwon, D.S., LaMar, M., Lampiris, H., Lanzafame, M., Lederman, M.M., Lee, D.M., Lee, J.M.L., Lee, M.J., Lee, E.T.Y., Lemoine, J., Levy, J.A., Llibre, J.M., Liguori, M.A., Little, S.J., Liu, A.Y., Lopez, A.J., Loutfy, M.R., Loy, D., Mohammed, D.Y., Man, A., Mansour, M.K., Marconi, V.C., Markowitz, M., Marques, R., Martin, J.N., Martin, H.L., Mayer, K.H., McElrath, M.J., McGhee, T.A., McGovern, B.H., McGowan, K., McIntyre, D., Mcleod, G.X., Menezes, P., Mesa, G., Metroka, C.E., Meyer-Olson, D., Miller, A.O., Montgomery, K., Mounzer, K.C., Nagami, E.H., Nagin, I., Nahass, R.G., Nelson, M.O., Nielsen, C., Norene, D.L., O'Connor, D.H., Ojikutu, B.O., Okulicz, J., Oladehin, O.O., Oldfield, E.C., Olender, S.A., Ostrowski, M., Owen, W.F., Pae, E., Parsonnet, J., Pavlatos, A.M., Perlmutter, A.M., Pierce, M.N., Pincus, J.M., Pisani, L., Price, L.J., Proia, L., Prokesch, R.C., Pujet, H.C., Ramgopal, M., Rathod, A., Rausch, M., Ravishankar, J., Rhame, F.S., Richards, C.S., Richman, D.D., Rodes, B., Rodriguez, M., Rose, R.C., Rosenberg, E.S., Rosenthal, D., Ross, P.E., Rubin, D.S., Rumbaugh, E., Saenz, L., Salvaggio, M.R., Sanchez, W.C., Sanjana, V.M., Santiago, S., Schmidt, W., Schuitemaker, H., Sestak, P.M., Shalit, P., Shay, W., Shirvani, V.N., Silebi, V.I., Sizemore, J.M., Skolnik, P.R., Sokol-Anderson, M., Sosman, J.M., Stabile, P., Stapleton, J.T., Starrett, S., Stein, F., Stellbrink, H.-J., Sterman, F.L., Stone, V.E., Stone, D.R., Tambussi, G., Taplitz, R.A., Tedaldi, E.M., Telenti, A., Theisen, W., Torres, R., Tosiello, L., Tremblay, C., Tribble, M.A., Trinh, P.D., Tsao, A., Ueda, P., Vaccaro, A., Valadas, E., Vanig, T.J., Vecino, I., Vega, V.M., Veikley, W., Wade, B.H., Walworth, C., Wanidworanun, C., Ward, D.J., Warner, D.A., Weber, R.D., Webster, D., Weis, S., Wheeler, D.A., White, D.J., Wilkins, E., Winston, A., Wlodaver, C.G., van't Wout, A., Wright, D.P., Yang, O.O.,

Yurdin, D.L., Zabukovic, B.W., Zachary, K.C., Zeeman, B., and Zhao, M. (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science 330, 1551–1557.

Jaffe, A.E., and Irizarry, R.A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 15, R31.

Jäger, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li, K., Hernandez, H., Jang, G.M., Roth, S.L., Akiva, E., Marlett, J., Stephens, M., D'Orso, I., Fernandes, J., Fahey, M., Mahon, C., O'Donoghue, A.J., Todorovic, A., Morris, J.H., Maltby, D.A., Alber, T., Cagney, G., Bushman, F.D., Young, J.A., Chanda, S.K., Sundquist, W.I., Kortemme, T., Hernandez, R.D., Craik, C.S., Burlingame, A., Sali, A., Frankel, A.D., and Krogan, N.J. (2012). Global landscape of HIV-human protein complexes. Nature 481, 365–370.

Joshi, D., O'Grady, J., Dieterich, D., Gazzard, B., and Agarwal, K. (2011). Increasing burden of liver disease in patients with HIV infection. The Lancet 377, 1198–1209.

Kennedy, B.K., Berger, S.L., Brunet, A., Campisi, J., Cuervo, A.M., Epel, E.S., Franceschi, C., Lithgow, G.J., Morimoto, R.I., Pessin, J.E., Rando, T.A., Richardson, A., Schadt, E.E., Wyss-Coray, T., and Sierra, F. (2014). Geroscience: linking aging to chronic disease. Cell 159, 709–713.

König, R., Zhou, Y., Elleder, D., Diamond, T.L., Bonamy, G.M.C., Irelan, J.T., Chiang, C., Tu, B.P., De Jesus, P.D., Lilley, C.E., Seidel, S., Opaluch, A.M., Caldwell, J.S., Weitzman, M.D., Kuhen, K.L., Bandyopadhyay, S., Ideker, T., Orth, A.P., Miraglia, L.J., Bushman, F.D., Young, J.A., and Chanda, S.K. (2008). Global Analysis of Host-Pathogen Interactions that Regulate Early-Stage HIV-1 Replication. Cell 135, 49–60.

Kovari, H., Sabin, C.A., Ledergerber, B., Ryom, L., Worm, S.W., Smith, C., Phillips, A., Reiss, P., Fontas, E., Petoumenos, K., De Wit, S., Morlat, P., Lundgren, J.D., and Weber, R. (2013). Antiretroviral drug-related liver mortality among HIV-positive persons in the absence of hepatitis B or C virus coinfection: the data collection on adverse events of anti-HIV drugs study. Clin. Infect. Dis. 56, 870–879.

Kulkarni, S., Savan, R., Qi, Y., Gao, X., Yuki, Y., Bass, S.E., Martin, M.P., Hunt, P., Deeks, S.G., Telenti, A., Pereyra, F., Goldstein, D., Wolinsky, S., Walker, B., Young, H.A., and Carrington, M. (2011). Differential microRNA regulation of HLA-C expression and its association with HIV control. Nature 472, 495–498.

Leeansyah, E., Cameron, P.U., Solomon, A., Tennakoon, S., Velayudham, P., Gouillou, M., Spelman, T., Hearps, A., Fairley, C., Smit, D.V., Pierce, A.B.,

Armishaw, J., Crowe, S.M., Cooper, D.A., Koelsch, K.K., Liu, J.-P., Chuah, J., and Lewin, S.R. (2013). Inhibition of telomerase activity by human immunodeficiency virus (HIV) nucleos(t)ide reverse transcriptase inhibitors: a potential factor contributing to HIV-associated accelerated aging. J. Infect. Dis. 207, 1157–1165.

Leung, V., Gillis, J., Raboud, J., Cooper, C., Hogg, R.S., Loutfy, M.R., Machouf, N., Montaner, J.S.G., Rourke, S.B., Tsoukas, C., Klein, M.B., and CANOC Collaboration (2013). Predictors of CD4:CD8 ratio normalization and its effect on health outcomes in the era of combination antiretroviral therapy. PLoS ONE 8, e77665.

Lindsey, J., McGill, N.I., Lindsey, L.A., Green, D.K., and Cooke, H.J. (1991). In vivo loss of telomeric repeats with age in humans. Mutation Research/DNAging 256, 45–48.

Luz, P.M., Grinsztejn, B., Velasque, L., Pacheco, A.G., Veloso, V.G., Moore, R.D., and Struchiner, C.J. (2014). Long-term CD4+ cell count in response to combination antiretroviral therapy. PLoS ONE 9, e93039.

Maartens, G., Celum, C., and Lewin, S.R. (2014). HIV infection: epidemiology, pathogenesis, treatment, and prevention. Lancet 384, 258–271.

Maegawa, S., Hinkal, G., Kim, H.S., Shen, L., Zhang, L., Zhang, J., Zhang, N., Liang, S., Donehower, L.A., and Issa, J.-P.J. (2010). Widespread and tissue specific age-related DNA methylation changes in mice. Genome Res. 20, 332–340.

Marabita, F., Almgren, M., Lindholm, M.E., Ruhrmann, S., Fagerström-Billai, F., Jagodic, M., Sundberg, C.J., Ekström, T.J., Teschendorff, A.E., Tegnér, J., and Gomez-Cabrero, D. (2013). An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics 8, 333–346.

Marioni, R.E., Shah, S., McRae, A.F., Chen, B.H., Colicino, E., Harris, S.E., Gibson, J., Henders, A.K., Redmond, P., Cox, S.R., Pattie, A., Corley, J., Murphy, L., Martin, N.G., Montgomery, G.W., Feinberg, A.P., Fallin, M.D., Multhaup, M.L., Jaffe, A.E., Joehanes, R., Schwartz, J., Just, A.C., Lunetta, K.L., Murabito, J.M., Starr, J.M., Horvath, S., Baccarelli, A.A., Levy, D., Visscher, P.M., Wray, N.R., and Deary, I.J. (2015). DNA methylation age of blood predicts all-cause mortality in later life. Genome Biology 16.

Martin, M.P., and Carrington, M. (2013). Immunogenetics of HIV disease. Immunol. Rev. 254, 245–264.

McMichael, A.J., and Rowland-Jones, S.L. (2001). Cellular immune responses to HIV. Nature 410, 980–987.

Mills, E.J., Bärnighausen, T., and Negin, J. (2012). HIV and aging--preparing for the challenges ahead. N. Engl. J. Med. 366, 1270–1273.

Mocroft, A., Phillips, A.N., Gatell, J., Ledergerber, B., Fisher, M., Clumeck, N., Losso, M., Lazzarin, A., Fatkenheuer, G., Lundgren, J.D., and EuroSIDA study group (2007). Normalisation of CD4 counts in patients with HIV-1 infection and maximum virological suppression who are taking combination antiretroviral therapy: an observational cohort study. Lancet 370, 407–413.

Nightingale, S., Winston, A., Letendre, S., Michael, B.D., McArthur, J.C., Khoo, S., and Solomon, T. (2014). Controversies in HIV-associated neurocognitive disorders. Lancet Neurol 13, 1139–1151.

Pathai, S., Lawn, S.D., Gilbert, C.E., McGuinness, D., McGlynn, L., Weiss, H.A., Port, J., Christ, T., Barclay, K., Wood, R., Bekker, L.-G., and Shiels, P.G. (2013). Accelerated biological ageing in HIV-infected individuals in South Africa: a case-control study. AIDS 27, 2375–2384.

Rakyan, V.K., Down, T.A., Maslau, S., Andrew, T., Yang, T.-P., Beyan, H., Whittaker, P., McCann, O.T., Finer, S., Valdes, A.M., Leslie, R.D., Deloukas, P., and Spector, T.D. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 20, 434–439.

Reynolds, L.M., Taylor, J.R., Ding, J., Lohman, K., Johnson, C., Siscovick, D., Burke, G., Post, W., Shea, S., Jacobs Jr., D.R., Stunnenberg, H., Kritchevsky, S.B., Hoeschele, I., McCall, C.E., Herrington, D.M., Tracy, R.P., and Liu, Y. (2014). Age-related variations in the methylome associated with gene expression in human monocytes and T cells. Nature Communications 5, 5366.

Riboli, E., Hunt, K.J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondière, U.R., Hémon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D., Trichopoulou, A., Vineis, P., Palli, D., Bueno-De-Mesquita, H.B., Peeters, P.H.M., Lund, E., Engeset, D., González, C.A., Barricarte, A., Berglund, G., Hallmans, G., Day, N.E., Key, T.J., Kaaks, R., and Saracci, R. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr 5, 1113–1124.

Rickabaugh, T.M., Baxter, R.M., Sehl, M., Sinsheimer, J.S., Hultin, P.M., Hultin, L.E., Quach, A., Martínez-Maza, O., Horvath, S., Vilain, E., and Jamieson, B.D. (2015). Acceleration of Age-Associated Methylation Patterns in HIV-1-Infected Adults. PLOS ONE 10, e0119201.

Rickabaugh, T.M., Kilpatrick, R.D., Hultin, L.E., Hultin, P.M., Hausner, M.A., Sugar, C.A., Althoff, K.N., Margolick, J.B., Rinaldo, C.R., Detels, R., Phair, J., Effros, R.B.,

and Jamieson, B.D. (2011). The Dual Impact of HIV-1 Infection and Aging on Naïve CD4+ T-Cells: Additive and Distinct Patterns of Impairment. PLoS ONE 6, e16459.

Ruelas, D.S., and Greene, W.C. (2013). An integrated overview of HIV-1 latency. Cell 155, 519–529.

Serrano-Villar, S., Pérez-Elías, M.J., Dronda, F., Casado, J.L., Moreno, A., Royuela, A., Pérez-Molina, J.A., Sainz, T., Navas, E., Hermida, J.M., Quereda, C., and Moreno, S. (2014). Increased risk of serious non-AIDS-related events in HIV-infected subjects on antiretroviral therapy associated with a low CD4/CD8 ratio. PLoS ONE 9, e85798.

Serrano-Villar, S., Sainz, T., Lee, S.A., Hunt, P.W., Sinclair, E., Shacklett, B.L., Ferre, A.L., Hayes, T.L., Somsouk, M., Hsue, P.Y., Van Natta, M.L., Meinert, C.L., Lederman, M.M., Hatano, H., Jain, V., Huang, Y., Hecht, F.M., Martin, J.N., McCune, J.M., Moreno, S., and Deeks, S.G. (2014). HIV-infected individuals with low CD4/CD8 ratio despite effective antiretroviral therapy exhibit altered T cell subsets, heightened CD8+ T cell activation, and increased risk of non-AIDS morbidity and mortality. PLoS Pathog. 10, e1004078.

Shih, A.H., Abdel-Wahab, O., Patel, J.P., and Levine, R.L. (2012). The role of mutations in epigenetic regulators in myeloid malignancies. Nat. Rev. Cancer 12, 599–612.

Siebold, A.P., Banerjee, R., Tie, F., Kiss, D.L., Moskowitz, J., and Harte, P.J. (2010). Polycomb Repressive Complex 2 and Trithorax modulate Drosophila longevity and stress resistance. Proc. Natl. Acad. Sci. U.S.A. 107, 169–174.

Smith, R.L., de Boer, R., Brul, S., Budovskaya, Y., and van Spek, H. (2012). Premature and accelerated aging: HIV or HAART? Front Genet 3, 328.

Solomon, A., Tennakoon, S., Leeansyah, E., Arribas, J., Hill, A., Van Delft, Y., Moecklinghoff, C., and Lewin, S.R. (2014). No difference in the rate of change in telomere length or telomerase activity in HIV-infected patients after three years of darunavir/ritonavir with and without nucleoside analogues in the MONET trial. PLoS ONE 9, e109718.

Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 29, 189–196.

Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Weisenberger, D.J., Shen, H., Campan, M., Noushmehr, H., Bell, C.G., Maxwell, A.P., Savage, D.A., Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S.A., Jones, A., Beck, S., Wagner, W., Laird, P.W., Jacobs, I.J., and Widschwendter, M. (2010). Age-

dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Research 20, 440–446.

Thomas, R., Apps, R., Qi, Y., Gao, X., Male, V., O'hUigin, C., O'Connor, G., Ge, D., Fellay, J., Martin, J.N., Margolick, J., Goedert, J.J., Buchbinder, S., Kirk, G.D., Martin, M.P., Telenti, A., Deeks, S.G., Walker, B.D., Goldstein, D., McVicar, D.W., Moffett, A., and Carrington, M. (2009). HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. Nat. Genet. 41, 1290–1294.

Torres, R.A., and Lewis, W. (2014). Aging and HIV/AIDS: pathogenetic role of therapeutic side effects. Laboratory Investigation 94, 120–128.

Trono, D., Van Lint, C., Rouzioux, C., Verdin, E., Barré-Sinoussi, F., Chun, T.-W., and Chomont, N. (2010). HIV persistence and the prospect of long-term drug-free remissions for HIV-infected individuals. Science 329, 174–180.

Weidner, C., Lin, Q., Koch, C., Eisele, L., Beier, F., Ziegler, P., Bauerschlag, D., Jöckel, K.-H., Erbel, R., Mühleisen, T., Zenke, M., Brümmendorf, T., and Wagner, W. (2014). Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biology 15, R24.

West, J., Beck, S., Wang, X., and Teschendorff, A.E. (2013). An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. Sci Rep 3, 1630.

Zanet, D.L., Thorne, A., Singer, J., Maan, E.J., Sattha, B., Le Campion, A., Soudeyns, H., Pick, N., Murray, M., Money, D.M., Cote, H.C.F., and for the CIHR Emerging Team Grant on HIV Therapy and Aging: CARMA (2014). Association Between Short Leukocyte Telomere Length and HIV Infection in a Cohort Study: No Evidence of a Relationship With Antiretroviral Therapy. Clinical Infectious Diseases 58, 1322–1332.

Zhou, H., Xu, M., Huang, Q., Gates, A.T., Zhang, X.D., Castle, J.C., Stec, E., Ferrer, M., Strulovici, B., Hazuda, D.J., and Espeseth, A.S. (2008). Genome-scale RNAi screen for host factors required for HIV replication. Cell Host Microbe 4, 495–504.

(2006). CD4+ Count–Guided Interruption of Antiretroviral Treatment. New England Journal of Medicine 355, 2283–2296.